Notas de Análise de Sobrevivência

Uma Introdução com R

Pedro Serranho

Secção de Matemática
Departamento de Ciências e Tecnologia
Universidade Aberta

2015

Conteúdo

Capítulo 1

Introdução e Conceitos Base

A **Análise de Sobrevivência** é uma área da Estatística que adquiriu a sua designação devido à sua forte aplicação a estudos de sobrevivência de sujeitos a uma determinada doença ou causa de morte. De uma forma mais geral a Análise de Sobrevivência é a área que se debruça sobre estudos de tempos até à ocorrência de determinado evento de interesse. Como já foi dito, a área de aplicação é geralmente a sobrevivência a determinada doença, mas existem atualmente outras aplicações, desde o tempo de vida de componentes elétricos até ao tempo de incumprimento de crédito bancário, por exemplo.

Desta forma, na Análise de Sobrevivência a variável de interesse é contínua, uma vez que é o tempo até à ocorrência do evento de interesse. O evento de interesse é geralmente denominado neste contexto por **falha**, sendo que a variável de interesse é portanto geralmente denominada por **tempo de sobrevivência** . Esta pode ser dada em minutos, horas, dias, meses ou anos e como exemplo, o tempo de sobrevivência pode ser, entre outros, o tempo:

- de sobrevivência a determinada doença;
- até à cura de determinada doença;
- da cura até à recidiva, em especial em estudos de cancro;
- até à alta hospitalar, para determinada doença;
- de vida de um determinado componente elétrico;
- até ao incumprimento após crédito bancário;

Exercício 1.1. Pesquise alguns exemplos específicos de contextos para estudos de Análise de Sobrevivência e partilhe-os nos fóruns de turma.

Assim, o objetivo principal da análise de sobrevivência é estimar grandezas associadas com o tempo de sobrevivência, muitas delas baseadas da Função de Sobrevivência, que definimos de seguida. A função de sobrevivência indica a probabilidade de sobrevivência ao longo do tempo, pelo que a partir dela se pode sumariar os seguintes como os objetivos principais da análise de sobrevivência:

- Estimar a função de sobrevivência a partir de dados recolhidos;
- Comparar funções de sobrevivência entre grupos (a uma menor função de sobrevivência corresponde uma maior mortalidade);
- Estudar a influência de covariáveis de interesse na função de sobrevivência (por exemplo, determinar se o aumento do valor de determinada covariável faz aumentar ou diminuir a função de sobrevivência no intervalo de estudo).

Como veremos na secção seguinte a partir da função de sobrevivência consegue-se caracterizar a forma como a população sobrevive à doença ao longo do tempo.

1.1 Função de Sobrevivência

A **Função de Sobrevivência** é geralmente denotada por $S(t)$, em que t é o tempo. A função de sobrevivência no instante t indica a probabilidade da falha ainda não ter ocorrido, ou seja, a percentagem de sujeitos na população que ainda sobrevivem no instante t. Assim, sendo T a variável aleatória tempo de sobrevivência[1], temos a definição de função de sobrevivência dada por

$$S(t) = P(T \geq t).$$

É fácil verificar que a função de sobrevivência define a distribuição de probabilidade para a variável aleatória tempo de sobrevivência T. Na realidade, a função distribuição de probabilidade de T é definida por

$$F(t) = P(T \leq t),$$

logo esta é dada por

$$F(t) = 1 - S(t). \tag{1.1}$$

[1]Como habitualmente, consideramos T maiúsculo para designar a variável aleatória e t minúsculo para designar um determinado valor dessa variável.

Da mesma forma, no caso de funções suaves e bem comportadas, como a função densidade de probabilidade f satisfaz por definição

$$F(t) = \int_{-\infty}^{t} f(u)du$$

que no caso da variável tempo T (por assumir apenas valores positivos) se reduz a

$$F(t) = \int_{0}^{t} f(u)du,$$

temos, diferenciando em t de ambos os lados, que

$$F'(t) = f(t),$$

ou seja,

$$f(t) = -S'(t). \tag{1.2}$$

A função de sobrevivência permite também tirar conclusões sobre a sobrevivência a determinada doença ao longo do tempo. Permite, por exemplo, determinar algumas medidas de tendência central para o tempo de vida, como o **tempo de vida mediano** designado por t_{mediano}, que é definido pelo instante de tempo em que a função de sobrevivência é igual a 50%, ou seja,

$$S\left(t_{\text{mediano}}\right) = 0.5. \tag{1.3}$$

Temos também o **tempo de vida médio** denotado por t_m dado a partir da função de sobrevivência por

$$t_m = \int_{0}^{\infty} S(t)\, dt. \tag{1.4}$$

A função de sobrevivência permite também definir grandezas como a vida média residual a partir do instante t, denotada por $\mathrm{vmr(t)}$, que é definida como o tempo de vida médio restante para os sujeitos com idade t. Como a função de sobrevivência para sujeitos com idade t (ou seja, tornado t o instante inicial) é dada por $S(u)/S(t)$ para $u \geq t$, o tempo médio de vida a partir do instante t é dado por $\int_t^{\infty} \frac{S(u)}{S(t)} du$, e logo temos que a vida média residual definida por

$$\mathrm{vmr(t)} = \frac{1}{\mathrm{S(t)}} \int_{\mathrm{t}}^{\infty} \mathrm{S(u)du}. \tag{1.5}$$

É fácil verificar que $\mathrm{vmr}(0) = \mathrm{t_m}$. A função de sobrevivência está também fortemente relacionada com a taxa de falha, que explanaremos de seguida.

1.2 Taxa de Falha

A taxa de variação da falha num intervalo de tempo é a razão entre a probabilidade de uma falha ocorrer nesse intervalo de tempo e a amplitude do intervalo, condicionada ao facto da falha não ter acontecido antes. Considerando um instante t, a **taxa de falha** $\lambda(t)$ nesse instante é obtida fazendo o limite do intervalo anterior em torno de t para uma amplitude nula, ou seja, é definida por

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T \leq t+h | t \leq T)}{h}$$

que pode ser reescrita como

$$\lambda(t) = \lim_{h \to 0} \frac{S(t) - S(t+h)}{hS(t)}, \tag{1.6}$$

utilizando a definição da probabilidade condicionada e o facto de

$$\begin{aligned} P(t \leq T \leq t+h) &= P(T \leq t+h) - P(t \leq T) \\ &= F(t+h) - F(t) \\ &= (1 - S(t+h)) - (1 - S(t)) \\ &= S(t) - S(t+h). \end{aligned}$$

A taxa de falha é um indicador da variação instantânea da sobrevivência. Instantes com taxa de falhas maiores indicam maior decrescimento da função de sobrevivência do que instantes com taxas de falha menores. Da mesma forma, se a taxa de falha for crescente, isto indica que à medida que o tempo avança a probabilidade instantânea de morrer aumenta, o que traduz geralmente o efeito do envelhecimento. Além disso, a expressão (1.6) permite concluir que

$$\lambda(t) = \frac{1}{S(t)} \lim_{h \to 0} \frac{S(t) - S(t+h)}{h} = -\frac{1}{S(t)} \lim_{h \to 0} \frac{S(t+h) - S(t)}{h} = -\frac{S'(t)}{S(t)}. \tag{1.7}$$

Integrando ambos os lados da equação entre zero e t, obtemos

$$\int_0^t \lambda(u)\, du = -\left[\ln S(u)\right]_0^t,$$

e como $S(0) = 1$ (ou seja, no instante inicial todos os sujeitos estão vivos), temos $\ln(S(0)) = 0$ e logo

$$\int_0^t \lambda(u)\, du = -\ln S(t).$$

Desta relação entre a taxa de falha e a função de sobrevivência tiramos outras duas que são interessantes. Por um lado, aplicando a exponencial temos

$$S(t) = e^{-\int_0^t \lambda(u)\,du}. \tag{1.8}$$

Por outro diferenciando em t, temos

$$\lambda(t) = -(\ln S(t))'. \tag{1.9}$$

Pelas expressões anteriores, em especial pela expressão (1.8), vemos que a informação da taxa de falha chega muito suavizada à função de sobrevivência, por via da aplicação do integral, que tende a suavizar variações pontuais bruscas. Assim, a taxa de falha tende a dar maior informação que a função de sobrevivência, uma vez que funções de sobrevivência muito próximas podem ser originadas por taxas de falha muito distintas.

Exercício 1.2. Mostre que a taxa de falha é dada pela razão entre a função densidade de probabilidade da variável aleatória T e função de sobrevivência, isto é, que temos

$$\lambda(t) = \frac{f(t)}{S(t)}. \tag{1.10}$$

Resposta.

Como de (1.2) temos $f(t) = -S'(t)$, substituindo $S'(t)$ por f em (1.7), temos o resultado. ■

Uma outra função que aparece muitas vezes na literatura é a **taxa de falha acumulada**, que é definida por

$$\Lambda(t) = \int_0^t \lambda(u)\,du. \tag{1.11}$$

Esta função é mais fácil de ser estimada, pelo que, embora não tenha um significado direto, permite obter estimativa da taxa de falha a partir dela. É também fácil de perceber que a partir de (1.8) se obtem

$$S(t) = e^{-\Lambda(t)}. \tag{1.12}$$

ou em alternativa

$$\Lambda(t) = -\ln S(t). \tag{1.13}$$

Notamos que o conhecimento da função de sobrevivência S permite obter a taxa de falha λ através de (1.9), a função densidade de probabilidade f através

de (1.2), a função densidade de probabilidade (1.1), a taxa de falha acumulada a partir de (1.13), assim como os tempo médio (1.4) e mediano de vida (1.3) e a vida média residual (1.5). Assim, nos próximos capítulos dedicaremo-nos a formas de estimar a função de sobrevivência. Antes, no entanto, vamo-nos dedicar a alguns aspetos da montagem da recolha de dados e modelos, que são importantes no contexto de análise de sobrevivência.

1.3 Tempo de Sobrevivência

Para todas as grandezas e funções que temos considerado neste texto, é importante definir o tempo de sobrevivência de forma inequívoca para cada sujeito, de forma a que o tempo de sobrevivência tenha o significado de interesse para o estudo a ser feito. Neste texto consideraremos apenas um evento de interesse, isto é, a falha originada por uma causa de interesse. Por exemplo, para estudar a morte a determinado tipo de cancro, não se pode considerar uma falha a morte devido a outro factor como por exemplo um acidente de trabalho.

Além do instante de falha, para a definição do tempo de sobrevivência é de crucial importância a definição do instante de início de contagem desse tempo. Ao contrário dos estudos estatísticos usuais, geralmente o instante de início de contagem do tempo de sobrevivência não coincide com o início da recolha dos dados. Por exemplo, suponhamos que estudamos a sobrevivência a determinado cancro, utilizando os dados dos pacientes que se encontram inscritos em determinado instituto entre as datas X e Y. Supomos agora que um dos paciente falece (ou seja, ocorre uma falha) na data Z, que se encontra no período do estudo entre X e Y. Qual será então o tempo de sobrevivência a considerar? Ingenuamente o leitor poderia considerar que seria o tempo que decorre entre X (o início do estudo) e Z (a data de falha). Na realidade este tempo não tem significado estatístico para o estudo, pois a data de início de estudo X é pouco relevante para a progressão da doença no sujeito em causa. Assim, a data de início a considerar deverá, neste caso, ser a data de diagnóstico (ou, em alternativa, a data do início do tratamento), uma vez que o que tem relevância para o estudo é o tempo de sobrevivência após desta data. Dessa forma, para todos os sujeitos no estudo, o tempo de sobrevivência seria o tempo entre a data de diagnóstico e a data de morte (falha), que é de facto o tempo com interesse para o estudo de sobrevivência.

O exemplo anterior levanta outras questões que provavelmente já assolam a mente do leitor:

(a) O que fazer com indivíduos que no período de estudo não falecem, ou seja, não ocorre a falha?

(b) Como entrar no estudo com fatores que influenciam o tempo de sobrevivência?

Começamos pela questão (b), porque na realidade não a vamos responder para já. Apenas a mencionaremos na secção 1.5, sendo que a aprofundaremos mais adiante no texto, nomeadamente quando iniciarmos o estudo de estimativas para a função de sobrevivência. Num estudo de sobrevivência a cancro, obviamente que o estadio do cancro na data de diagnóstico, o género e a idade do sujeito (entre outros fatores) podem influenciar o tempo de sobrevivência. Dessa forma, é necessário considerar estes fatores como covariáveis no modelo, ou seja, como variáveis independentes que poderão influenciar a variável (dependente) de interesse: o tempo de sobrevivência. Trataremos desse aspeto a partir do próximo capítulo.

Quanto à questão (a), o leitor provavelmente poderia sugerir (mais uma vez de forma ingénua) que os dados relativos a sujeitos em que a falha não ocorreu poderiam simplesmente ser eliminados do estudo. No entanto, é fácil de ver que se perde informação se se proceder dessa forma. Se sabemos para determinado sujeito que a falha não ocorreu no período de observação, sabemos que o seu tempo de sobrevivência é superior a determinado tempo T, pelo que essa informação não deve ser desprezada no modelo. Esta é outra grande diferença da Análise de Sobrevivência em relação a outros estudos estatísticos: permite utilizar dados censurados, isto é, dados em que a falha não ocorre no período de observação.

1.4 Censura e Truncamento

Entende-se por **censura** a observação parcial da resposta. Sempre que para determinado sujeito não ocorre a falha até ao tempo t de observação, dizemos que em t ocorreu uma censura. Em presença de censura, o que sabemos é que o tempo de sobrevivência é superior a t, tratando-se neste caso de uma **censura à direita**. A censura à direita pode ter várias origens, como por exemplo, a saída de um sujeito do estudo por ter mudado a região de residência, a cura do paciente ou a morte devida a outra causa, que não a doença em estudo. Existem contextos no entanto em que temos **censura à esquerda**, isto é, em que quando a censura ocorre sabemos que o tempo de sobrevivência é inferior ao censurado. Por exemplo, o caso em que se estuda em determinada localidade em que idade

as crianças começam a caminhar. Podem existir crianças que já sabem caminhar e em que não se saiba a data precisa em que isso aconteceu, pelo nesse caso a censura é à esquerda.

Assim, cada i-ésimo sujeito é caracterizado pelo tempo t_i e por uma variável dicotómica δ_i que assume os valores

$$\delta_i = \begin{cases} 1, & \text{em } t_i \text{ ocorre uma falha} \\ 0, & \text{em } t_i \text{ ocorre uma censura} \end{cases} \tag{1.14}$$

Desta forma, o conjunto de dados base em análise de sobrevivência são pares $(t_i, \delta_i), \quad i = 1, 2, \ldots, n$, que formam o conjunto de n pacientes observados, em que em alguns ocorreram falhas e noutros censuras.

Um outro fenómeno que é geralmente confundido com a censura é o **truncamento** , embora seja um conceito totalmente diferente. O truncamento é o fenómeno que impede indivíduos de fazerem parte da amostra e logo do estudo. Por exemplo, se para estudar a taxa de sobrevivência a determinada doença em determinada localidade eu recorrer aos dados do recenseamento eleitoral na junta de freguesia, apenas considero os sujeitos maiores de idade. Dessa forma, existe uma truncatura à esquerda, uma vez que os menores de idade são excluídos do estudo.

1.5 Covariáveis

Juntamente com o par (T_i, δ_i) definido anteriormente, podem (e devem) ser recolhidos para cada sujeito os valores de outras covariáveis de interesse, que se espera possam influenciar o tempo de sobrevivência. Ao incluir covariáveis, obtemos diferenças na função de sobrevivência consoante o valor da covariável e assim podemos testar se existem diferenças significativas entre valores diferente e grupos de interesse. No contexto de estudos de sobrevivência a doenças, salientamos as seguintes covariáveis, que têm geralmente um efeito significativo e que portanto devem ser consideradas:

- **Coorte**: Em estudos de sobrevivência a doenças, a função de sobrevivência é fortemente afetada pela época temporal em que é feito o estudo. É evidente que a taxa de sobrevivência ao cancro tem vindo a aumentar ao longo dos anos, devido aos avanços da medicina. Assim, caso tenhamos dados transversais a várias épocas, é importante introduzir no modelo uma variável coorte. Por exemplo, se tivermos dados de sobrevivência a cancro recolhidos em períodos de observação entre 1981 e 1984, depois entre 1990

e 1992 e entre 1996e 1999, será indicado considerar no modelo uma variável coorte com valores 0, 1 e 2, consoante o período em que o sujeito foi observado, uma vez que é esperado que a função de sobrevivência seja diferente consoante o coorte considerado. Os Coortes podem ser décadas, séculos ou períodos definidos no tempo e permitem ao modelo dar estimativas diferentes consoante o coorte considerado.

- **Grupo de Controlo vs. Grupo de Estudo:** A análise de sobrevivência tem aplicações para mostrar que determinado(s) tratamento(s) inovador(es) tem melhores resultados que o tratamento convencional. Assim, é importante recolher dados de sujeitos em vários grupos diferentes, como o grupo de controlo (sujeitos sem tratamento tomando um placebo, ou com o tratamento convencional) e o grupo de estudo (sujeito ao tratamento A, B e/ou C). Assim, a variável Grupo deve ser considerada no modelo com valores 0 para o grupo de controlo e 1,2,3,... para os grupos sujeitos ao tratamento A, B, C, etc...

- **Outras covariáveis:** Consoante a doença e o objetivo do estudo importa recolher dados de outros fatores que possam influenciar o tempo de sobrevivência. Por exemplo, em geral, a idade do paciente na data de diagnóstico é importante para o modelo. A idade do sujeito influencia a progressão da maioria das doenças. O género do sujeito é uma variável de interesse na maioria dos casos, uma vez que influencia não só a taxa de incidência, mas também a progressão da doença. Em estudos de cancro, os estadio do cancro no momento do diagnóstico influencia fortemente a função de sobrevivência, pelo que o estadio deve ser considerado no modelo.

Estamos agora em condições de iniciarmos o estudo de métodos para a estimativa da função de sobrevivência, assunto ao qual nos dedicaremos nos próximos capítulos.

Capítulo 2

Estimadores Não paramétricos

Os estimadores não paramétricos não assumem qualquer modelo de distribuição para a associação entre as covariáveis de interesse e a taxa de falha acumulada. Assim, estes são aplicáveis na maioria dos casos em que não existe informação para o modelo próprio da função de sobrevivência.

De notar que os estimadores têm de ter a capacidade de lidar com censuras, pois de outra forma não seriam aplicáveis ao contexto de Análise de Sobrevivência. No entanto, e como motivação, vamos começar por considerar o caso sem censuras, isto é, o caso em que para todos os sujeitos ocorreu uma falha no período de observação.

Exemplo 2.1. Suponhamos que recolhemos os dados de falhas ocorridas nos vários instantes de tempo tabelados abaixo, para o total de 168 sujeitos:

t_i	Nº de falhas	Nº de sobreviventes
0	-	168
1	5	163
2	7	156
3	10	146
4	25	121
5	52	69
6	35	34
7	19	15
8	15	0

Neste caso sem censuras, temos então como possibilidade para estimativa

para a função de sobrevivência

$$\hat{S}(t_i^+) = \frac{\text{nº de sujeitos vivos no instante } t_i}{\text{nº de sujeitos inicialmente no estudo}} \tag{2.1}$$

em que a notação t^+ indica que a estimativa é válida entre o instante t_i e o seguinte t_{i+1}. Por exemplo, para $t = 3$ teríamos a estimativa

$$\hat{S}(3) = \frac{136}{168} = 0.86905$$

ou seja, cerca de 86.9% dos sujeitos sobrevive ao instante $t = 3$. Da mesma forma, se pretendessemos estimar a taxa de falha, teríamos, por exemplo a estimativa

$$\hat{\lambda}([t_i, t_{i+1}[) = \frac{\text{nº de falhas entre } t_i \text{ e } t_{i+1}}{\text{nº de sujeitos vivos no instante } t_i}$$

pelo que para o intervalo $[3, 4[$ teríamos

$$\hat{\lambda}([3, 4[) = \frac{25}{146} = 0.17123.$$

Completando a tabela, teríamos as estimativas

t_i	Nº de falhas	Nº de sobreviventes	$\hat{S}(t_i^+)$	$\hat{\lambda}([t_i, t_{i+1}[)$
0	—	168	1	0.02976
1	5	163	0.97024	0.04294
2	7	156	0.92857	0.06410
3	10	146	0.86905	0.17123
4	25	121	0.72024	0.42975
5	52	69	0.41071	0.50725
6	35	34	0.20238	0.55882
7	19	15	0.08929	1
8	15	0	0	—

De notar que quando não há censuras, a função de sobrevivência é nula após o último instante medido, uma vez que para todos os sujeitos existe uma falha no período observado. Mais ainda, a taxa de falha é 1 no último intervalo medido, uma vez que todos os que ainda sobreviviam acabam por ter uma falha nesse período, devido à razão anterior. No entanto, este é problema no contexto geral de análise de sobrevivência, uma vez que teremos de ter estimativas que possam permitir a sobrevivência de alguns sujeitos, o que não acontece no caso das censuras. Assim fica explícito que a eliminação das censuras para um estudo de Análise de sobrevivência desvirtua totalmente os resultados.

Assim, iniciamos o estudo de técnicas que permitam considerar censuras, começando pelo estimador de Kaplan-Meier.

2.1 Estimador de Kaplan-Meier

O **estimador de Kaplan-Meier** foi sugerido por Kaplan e Meier em 1958 [1]. Na realidade, o estimador de Kaplan-Meier é a estimativa de máxima verosimilhança da função de sobrevivência, mas pode ser visto como uma adaptação da estimativa empírica considerada em (2.1), sendo que para a contabilidade do número de sujeitos vivos no instante t_i são contabilizados também os sujeitos que tiveram uma censura num instante superior. Na realidade, o estimador de Kaplan-Meier é uma função em escada em que o início dos degraus da escada coincídem com os instantes onde ocorrem falhas. Assim, o primeiro passo para a estimativa de Kaplan-Meier é dividir o intervalo em n sub-intervalos, com limites nos pontos de falha. A partir daí, começamos por estimar a função de sobrevivência no primeiro instante de falha t_1 como a razão entre o número de sujeitos em risco no instante t_1 sobre o número se sujeitos inicialmente no estudo. No segundo instante de falha t_2 consideramos a relação

$$P(T \geq t_2) = P(T \geq t_1, T \geq t_2) = P(T \geq t_1).P(T \geq t_2 | t \geq 1)$$

ou seja, a probabilidade do tempo de sobrevivência ser superior ou igual a t_2 é igual à do tempo de sobrevivência ser superior ou igual a t_1 vezes a probabilidade do tempo de sobrevivência ser superior ou igual a t_2 condicionada ao tempo ser superior a t_1. Dito em linguagem corrente, para a probabilidade de um sujeito sobreviver até t_2 é igual à probabilidade de sobreviver até t_1 vezes a probabilidade de sobreviver até t_2 sabendo que sobreviveu até t_1. Este processo é depois repetido sucessivamente para se ter as estimativas sucessivas nos vários instantes de falha.

Vejamos um exemplo, antes de explicitarmos a expressão geral.

Exemplo 2.2. Consideramos os dados recolhidos e ordenados de tempos de falha seguintes, em que o sinal + indica uma censura em vez de uma falha:

$$1, 3+, 4, 5, 5+, 6+, 7+, 8, 8, 9+, 10, 11+, 13+ .$$

Começamos por estabelecer os extremos dos intervalos t_i a considerar para a estimativa de Kaplan-Meier, ou seja,

$$t_0 = 0 \text{ (instante inicial)}, \; t_1 = 1, \; t_2 = 4, \; t_3 = 5, \; t_4 = 8, \; t_5 = 10.$$

Em seguida, para cada intervalo $[t_i, t_{i+1}[$ é preciso definir o número n_i de sujeitos em risco (ou seja, vivos) e o número f_i de falhas no intervalo, dando origem à tabela seguinte:

t_i	Intervalo	n_i	f_i
0	[0,1[	13	0
1	[1,4[	13	1
4	[4,5[	11	1
5	[5,8[	10	1
8	[8,10[	6	2
10	[10, 13[	3	1

A probabilidade de sobrevivência inicial é 1, logo consideramos que

$$S(0^+) = P(T \geq 0) = 1,$$

ou seja, a probabilidade de sobreviver ao instante 0 é 1.

De seguida, temos que no instante $t = 1$ ocorre 1 falha em 13 sujeitos em risco. Assim, a estimativa para o sujeito não sobreviver ao instante $t = 1$ é 1/13 (número de falhas sobre número de sujeitos em risco), pelo que a estimativa para sobreviver ao instante 1 é dada por

$$\hat{S}(1^+) = (1 - 1/13) = 0.92308.$$

No instante seguinte $t = 4$, ocorre 1 falha em 11 sujeitos em risco. A estimativa para o sujeito não sobreviver até ao instante $t = 4$, condicionada ao sujeito estar vivo no instante anterior $t = 1$ é então dada por

$$P(T \leq 4|T \geq 1) = 1/11 = 0.090909,$$

pelo que

$$P(T \geq 4|T \geq 1) = 1 - 1/11 = 0.90909.$$

Assim, a estimativa da função de sobrevivência é dada por

$$\hat{S}(4^+) = \underbrace{P(T \geq 1)}_{\hat{S}(1^+)}.P(T \geq 4|T \geq 1) = 0.92308 \times 0.90909 = 0.83916.$$

De forma semelhante, a estimativa para a probabilidade de sobrevivência ao instante $t = 5$, condicionada ao sujeito estar vivo no instante $t = 4$ é

$$P(T \geq 5|T \geq 4) = 1 - 1/10 = 0.9$$

pelo que a estimativa para o sujeito estar vivo no instante $t = 5$ é dada por

$$\hat{S}(5^+) = \hat{S}(4^+).P(T \geq 5|T \geq 4) = 0.83916 \times 0.9 = 0.75524.$$

Repetindo o processo, temos a estimativa de Kaplan-Meier para a função de sobrevivência dada na tabela seguinte:

Tabela 2.1: Estimativa de Kaplan Meier para a função de sobrevivência do exemplo 2.2.

t_i	Intervalo	n_i	f_i	$\hat{S}(t_i^+)$
0	[0,1[	13	0	1
1	[1,4[	13	1	0.92308
4	[4,5[	11	1	0.83916
5	[5,8[	10	1	0.75524
8	[8,10[	6	2	0.50350
10	[10, 13[	3	1	0.33566

Exercício 2.3. Confirme os valores obtidos da estimativa de Kaplan-Meier na tabela, que não foram calculados no texto.

Temos então a definição do estimador de Kaplan-Meier seguinte, que em [1] foi demonstrado ser o estimador de máxima verosimilhança da função de sobrevivência S.

Definição 2.4 (Estimador de Kaplan-Meier)**.**
Sejam para $i = 1, 2, \ldots, n$

- t_i os instantes em que foram observadas falhas na amostra e
- n_i os número de sujeitos em risco no instante t_i;
- f_i o número de falhas ocorridas no intervalo $[t_i, t_{i+1}]$;

Então, o **estimador de Kaplan-Meier** para a função de sobrevivência é dado pela função em escada

$$\hat{S}(t) = \prod_{i:t \geq t_i} \left(1 - \frac{f_i}{n_i}\right). \tag{2.2}$$

Nota 2.5 (Normalidade e Consistência). Como nota, referimos que é possível mostrar (ver [2]) que a distribuição do estimador de Kaplan-Meier converge assintoticamente para uma distribuição normal e que este é fracamente consistente, ou seja, para qualquer $\varepsilon > 0$ quando a dimensão n da amostra tende para infinito, temos

$$\lim_{n \to \infty} P(|\hat{(}S) - S| < \varepsilon) = 1.$$

Felizmente, hoje é possível utilizar recursos computacionais para obter este tipo de estimativa. Ainda que a quantidade de dados utilizada no exemplo 2.2

para fins ilustrativos é relativamente pequena, em casos de dados recolhidos para estudos de clínicos a quantidade de dados a tratar torna inviável o recurso a cálculo manual para estimar a função de sobrevivência. Apresentamos o exemplo seguinte como forma de obter a mesma estimativa para o exemplo descrito, utilizando o software R.

Exercício R 2.6. Utilize o R para obter a estimativa de Kaplan-Meier para a função de sobrevivência para os dados do exemplo 2.2.

Resposta.

Após instalar o *package* `survival`, começamos por carregá-lo para a sessão de R em curso através do comando:

```
require(survival)
```

Este *package* contém todas as funções relevantes para este texto, no âmbito de estudos de análise de sobrevivência. De seguida criamos as listagens dos tempos e a lista censura que indica se o tempo recolhido corresponde a uma falha (1) ou censura (0):

```
t<-c(1,3,4,5,5,6,7,8,8,9,10,11,13)
cens<-c(1,0,1,1,0,0,0,1,1,0,1,0,0)
```

Com o comando seguinte, obtemos o modelo de Kaplan-Meier na variável `ekm`, através do uso da função `survfit` como indicado:

```
ekm<-survfit(Surv(t,cens)~1, conf.int=0)
```

O comando `summary` permite obter um sumário dos resultados, nomeadamente a tabela com a estimativa da estimativa de Kaplan-Meier que comprova o resultado obtido na tabela 2.1.

```
summary(ekm)
```

Finalmente, com o comando seguinte, obtemos o gráfico da função de sobrevivência estimada pelo estimador de Kapla-Meier, que apresentamos na figura 2.1.

```
plot(ekm,lty=c(2,1),xlab="Tempo",
        ylab="S(t) por Kaplan-Meier")
```

De notar que no gráfico da estimativa da função de sobrevivência, os instantes em que ocorrem censuras são assinalados automaticamente com o símbolo +. ■

De notar que como em qualquer estimador, é importante ter noção da sua variância, para assim estimar um intervalo de confiança adequado. Como o estimador tem uma distribuição assintótica normal (ver nota 2.5), esperamos que o

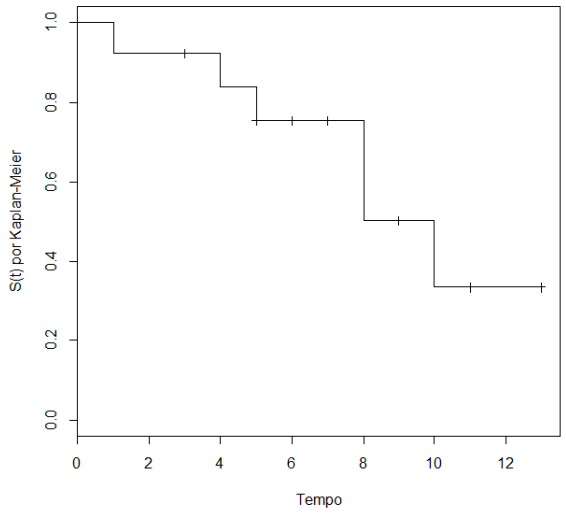


Figura 2.1: Estimativa de Kaplan Meier para a função de sobrevivência do exemplo 2.2.

intervalo de confiança (a $(1-\alpha)$ de confiança) seja dado em cada instante t por

$$IC_{1-\alpha}(t) = \left[\hat{S}(t) + z_{\alpha/2}\sqrt{\text{Var}\left(\hat{S}(t)\right)}, \hat{S}(t) + z_{1-\alpha/2}\sqrt{\text{Var}\left(\hat{S}(t)\right)}\right],$$

em que z_α é o percentil α da distribuição normal. Assim, falta obter a variância do estimador.

Teorema 2.7 (Fórmula de Greenwood).
A variância assintótica do estimador de máxima verosimilhança de Kaplan-Meier é dado pela fórmula de Greenwood

$$\widehat{\text{Var}}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{i:t\geq t_i} \frac{f_i}{n_i(ni - f_i)}. \tag{2.3}$$

Demonstração. A prova vem das propriedades do estimador de máxima verosimilhança e pode ser detalhada em [3].

Exercício R 2.8. Determine o Intervalo de Confiança a 95% segundo a fórmula de Greenwood para a função de sobrevivência do exemplo 2.2.

Resposta.

Deixamos o cálculo manual como exercício e passamos a ilustrar a forma de utilizar o R para o efeito. Em relação ao algoritmo 2.6, as diferenças são pequenas. Começamos por carregar o *package* `survival` e os dados, conforme no algoritmo 2.6. Depois pedimos o estimador de K-M, considerando o intervalo de confiança com fórmula de Greenwood com o comando

```
ekmICG<-survfit(Surv(t,cens)~1,conf.type="plain",
         conf.int=0.95)
```

apresentado o seu resultado através da função `summary`, ou seja,

```
summary(ekmICG)
```

De notar que como output temos o valor das estimativas para função de sobrevivência $\hat(S)(t)$, o erro padrão $\sqrt{\mathrm{Var}\left(\hat{S}(t)\right)}$ e os limites inferiores e superiores do intervalo de confiança a 95%. Estes podem ser sumariados na tabela seguinte

t_i	n_i	f_i	$\hat{S}(t_i^+)$	erro padrão	IC
1	13	1	0.923	0.0739	[0.778,1.000]
4	11	1	0.839	0.1045	[0.634,1.000]
5	10	1	0.755	0.1232	[0.514,0.997]
8	6	2	0.503	0.1669	[0.176,0.831]
10	3	1	0.336	0.1765	[0.000,0.682]

Para obter graficamente o intervalo de confiança (ver figura 2.2), basta fazer agora

```
plot(ekmICG,lty=c(1,2,2),xlab="Tempo",ylab="S(t) K-M+IC")
```

■

De notar que pela fórmula de Greenwood, o intervalo de confiança pode ter limites superiores a 1 ou inferiores a zero, o que no contexto de uma função de sobrevivência não faz sentido[1]. Desta forma, foi sugerida em 1980 uma transformação logarítmica da função de sobrevivência em [3] por Kalbfleish e Prentice, definindo uma nova variável

$$\hat{U}(t) = \ln\left(-\ln \hat{S}(t)\right).$$

[1] A função `survfit` do R elimina este aspeto apresentando os limites superiores a 1 no IC como 1 e os inferiores a zero como zero.

É possível mostrar que para o estimador $\hat{U}$ a variância é dada por

$$\widehat{\text{Var}}\left(\hat{U}(t)\right) = \frac{\sum_{i:t_i<t} \frac{f_i}{n_i(n_i-f_i)}}{\left(\ln \hat{S}(t)\right)^2}$$

e que assim, o intevalo de confiança para $S(t)$ é dado por

$$IC_{1-\alpha}(t) = \left[\hat{S}(t)^{\exp\left(z_{\alpha/2}\sqrt{\widehat{\text{Var}}\left(\hat{U}(t)\right)}\right)}, \hat{S}(t)^{\exp\left(z_{1-\alpha/2}\sqrt{\widehat{\text{Var}}\left(\hat{U}(t)\right)}\right)}\right].$$

Como é fácil de demonstrar, o intervalo anterior tem sempre limites entre 0 e 1, uma vez que a própria estimativa $\hat{S}(t)$ para a função de sobrevivência tem sempre valores entre 0 e 1.

Exercício R 2.9. Determine o Intervalo de Confiança a 95% segundo a transformação logarítmica de Kalbfleish e Prentice para a função de sobrevivência do exemplo 2.2.

Resposta.

Novamente deixamos o cálculo manual como exercício e passamos a ilustrar a forma de utilizar o R para o efeito. Depois de carregar o *package* `survival` e os dados, conforme no algoritmo 2.6, pedimos o estimador de K-M, considerando o intervalo de confiança com transformação logarítmica de Kalbfleish e Prentice com os comandos

```
ekmICln<-survfit(Surv(t,cens)~1,conf.type="log-log",
        conf.int=0.95)
summary(ekmICln)
plot(ekmICln,lty=c(1,2,2),xlab="Tempo",ylab="S(t) K-M+ICln")
```

Os resultados obtidos estão na figura 2.2 e são os seguintes:

t_i	n_i	f_i	$\hat{S}(t_i^+)$	erro padrão	IC
1	13	1	0.923	0.0739	[0.5664,0.989]
4	11	1	0.839	0.1045	[0.4940,0.957]
5	10	1	0.755	0.1232	[0.4161,0.914]
8	6	2	0.503	0.1669	[0.1705,0.766]
10	3	1	0.336	0.1765	[0.0604,0.654]

■

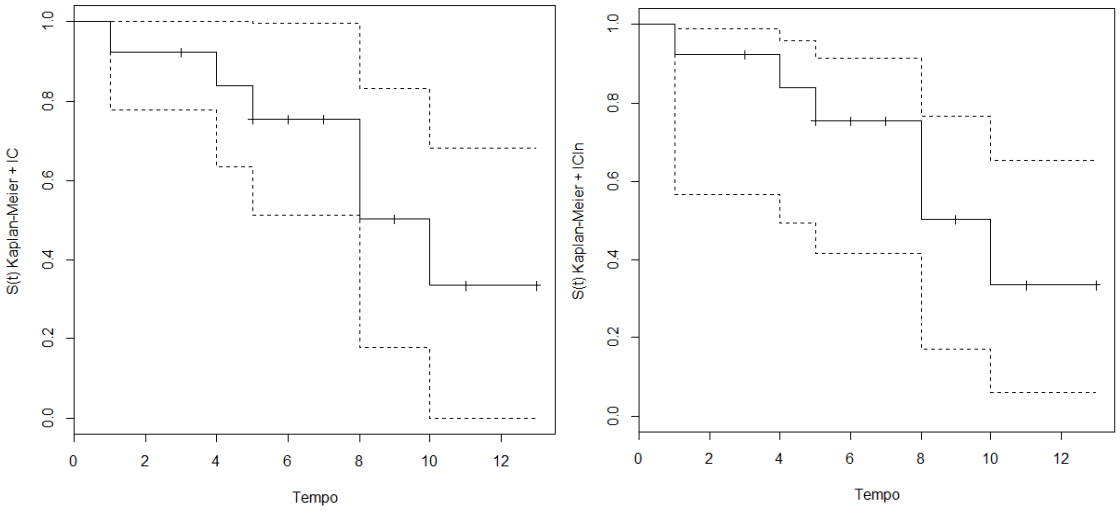


Figura 2.2: Estimativa de Kaplan Meier para a função de sobrevivência do exemplo 2.2 com intervalo de Confiança para a fórmula de Greenwood (à esquerda) e transformação logarítmica de Kalbfleish e Prentice (à direita).

2.2 Comparação entre grupos

Uma das aplicaçõe principais da análise de sobrevivência é a comparação entre grupos, isto é, determinar se existem diferenças estatisticamente significativas entre as funções de sobrevivência de dois ou mais grupos distintos da população. Os grupos podem ser várias regiões do país, várias faixas etárias, grupos de controlo e de tratamento, entre outros. Nesta secção veremos como podemos aplicar o uso de R para obter estimativas para diferentes grupos e como podemos testar se existem diferenças estatisticamente significativas.

Comecemos por considerar o seguinte exemplo, que utilizaremos para ilustrar as técnicas utilizadas nas seguintes linhas.

Exemplo 2.10. Sujeitos com determinada doença foram separados aleatoriamente em 3 grupos distintos. Ao grupo de controlo foi administrado um placebo. Ao grupo do tratamento 1, foi administrado um medicamento em duas doses diárias. Ao grupo de tratamento 2 foi administrado a mesma dosagem diária de medicamento mas em 3 doses diárias. Foi registado para cada um dos sujeitos o tempo em dias até obterem alta hospitalar. Consideramos os dados recolhidos e ordenados de tempos de falha seguintes, em que o sinal + indica uma censura em vez de uma falha:

Controlo	Tratamento 1	Tratamento 2
5, 8, 8+, 9, 10, 10+, 14	4, 5, 5+, 6, 8, 9, 9+, 10+, 11, 12+	5+, 5, 6, 6+, 7, 8, 8+, 10+

Exercício R 2.11. Determine a estimativa de Kaplan-Meier da função de sobrevivência para cada um dos três grupos do exemplo 2.10.

Resposta.

Tal como nos exemplos anteriores, necessitamos de introduzir os tempos e as censuras, o que pode ser feito com os comandos:

```
require(survival)
t<-c(5,8,8,9,10,10,14,4,5,5,6,8,9,9,10,11,12,5,5,6,6,7,8,8,10)
cens<-c(1,1,0,1,1,0,1,1,1,0,1,1,1,0,0,1,0,0,1,1,0,1,1,0,0)
```

De seguida, introduzimos (pela ordem corresponde aos tempos e censuras) a variável grupos, por exemplo, como

```
gr<-c(rep(0,7),rep(1,10),rep(2,8))
```

Temos então a estimativa de Kaplan-Meier para a função de sobrevivência em cada grupo e os respetivo gráfico dados pelos comandos:

```
ekm<-survfit(Surv(t,cens)~gr)
summary(ekm)
plot(ekm, lty=c(1,2,3), xlab="Tempo", ylab="S(t) por
        Kaplan-Meier")
legend(10,0.95,lty=c(1,2,3),c("Controlo","Tratamento 1",
        "Tratamento 2"))
```

Os resultados são apresentados no figura 2.3.

■

A comparação na figura 2.3 ilustra-nos que aparentemente não existe grande diferença entre os vários grupos. No entanto, é necessário fazer um teste estatístico adequado para a verificação desta hipótese, o que trataremos de seguida.

2.2.1 Teste Logrank

O teste de logrank permite testar a hipótese nula de comparação de funções de sobrevivência entre p grupos dada por

$$\begin{cases} S_1(t) = S_2(t) = \cdots = S_p(t), & \text{no intervalo de observação,} \\ S_i(t) \neq S_j(t) & \text{para algum par } (i,j), \text{ no intervalo de observação,} \end{cases}$$

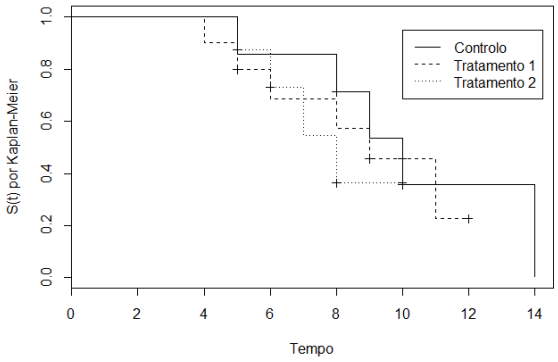


Figura 2.3: Estimativa de Kaplan Meier para a função de sobrevivência em cada grupo do exemplo 2.10.

Este teste é talvez o mais usado no contexto de análise de sobrevivência, uma vez que permite testar igualmente em todo o intervalo, em vez de fazer a comparação assintótica para t grande. No entanto, como hipótese para a aplicação do teste supõe-se que a razão entre as taxa de falha em cada grupo é sensivelmente constante ao longo do tempo, hipótese que também será necessária mais tarde para o modelo de Cox.

O teste de logrank parte da listagem ordenada de todos os tempos de falha em todos os grupos dada por

$$t_1 < t_2 < \cdots < t_m.$$

Em cada instante t_k, podemos construir a tabela seguinte, em que se assinalam por grupo i e no total o número de falhas f_{ik}, f_k e o número de não falhas dados por $n_{ik} - f_{ik}, n_k - f_k$, respetivamente, no instante t_k, em que n_{ik} é o número de sobreviventes no grupo i e n_k é o número total de sobreviventes no instante t_k:

	Grupo 1	Grupo 2	...	Grupo p	Total
Falha	f_{1k}	f_{2k}	...	f_{pk}	f_k
Não Falha	$n_{1k} - f_{1k}$	$n_{2k} - f_{2k}$	...	$n_{pk} - f_{pk}$	n_k-f_k
Total	n_{1k}	n_{2k}	...	n_{pk}	n_k

A partir da tabela de contingência anterior, podemos estabelecer o valor esperado para o número de falhas no grupo i no instante j no caso da hipótese nula (de funções de sobrevivência iguais), que seria dado por $e_{ik} = \frac{f_k n_{ik}}{n_k}$. Notando agora que a distribuição conjunta da variável $(f_{2k}, f_{3k}, \ldots, f_{pk})$ é uma hipergeométrica multivariada, consegue-se obter a estatística de teste

$$T = e' V^{-1} e \sim \chi^2_{p-1}$$

que tem distribuição qui-quadrado com $p-1$ graus de liberdade e em que V é a matriz de covariância das variáveis $(f_{2k}, f_{3k}, \ldots, f_{pk})$ e o vetor e é simplesmente a soma em todos os instantes k da diferença entre o valor observado e o valor esperado para o número de falhas dado por

$$e = \left(\sum_{k=1}^{m} f2k - e_{2k}, \sum_{k=1}^{m} f3k - e_{3k}, \ldots, \sum_{k=1}^{m} fpk - e_{pk} \right).$$

Exercício R 2.12. Determine se existe diferença significativas entre as funções de sobrevivência dos três grupos do exemplo 2.10.

Resposta.

Para aplicar o teste de logrank em R, após ter executado os comandos do algoritmo 2.11, basta executar o comando:

```
survdiff(Surv(t,cens)~gr,rho=0)
```

Como se verifica pelo output, temos um valor-p de 0.795, pelo que para uma significância de 5% não existe evidência estatística de diferenças entre os três grupos. Assim, não só não existe diferenças entre a administração do medicamento em 2 ou 3 doses, nem existe diferença entre a administração ou não do medicamento. Desta forma, concluímos que não existe evidência estatística que o efeito do medicamento seja eficaz[2]. ■

2.3 Estimador de Nelson-Aalen

Embora seja o mais utilizado, o estimador de Kaplan-Meier é apenas um entre muitos estimadores não-paramétricos disponíveis. Um desses estimadores é o estimador de Nelson-Aalen. Ao contrário do estimador de Kaplan-Meier que se baseia diretamente na estimação de sobrevivência, o estimador de Nelson-Aalen baseia-se na estimação da função de taxa acumulada, obtendo a estimativa da função de sobrevivência através da relação (1.12). Assim, partindo da definição de taxa de falha acumulada (1.11), o estimador de Nelson-Aalen [4, 5] para a função taxa de falha é dado por

$$\hat{\Lambda}(t) = \sum_{i:t<t_i} \frac{f_i}{n_i} \tag{2.4}$$

[2]De notar que a dimensão da amostra utilizada é muito pequena, pelo que é possível que a não significância seja apenas resultados deste aspeto, pelo que se deve ser prudente nas conclusões.

em que f_i é o número de falhas no instante t_i e n_i é o número de sujeitos em risco no instante i. Intuitivamente, a taxa de falha acumulada é dada pela soma das taxas de falhas verificadas até ao instante t, definidas por $\frac{f_i}{n_i}$. Consegue-se mostrar [5] que a variância assintótica deste estimador é dada por

$$\widehat{\mathrm{Var}}\left(\hat{\Lambda}\right)(t) = \sum_{i:t<t_i} \frac{f_i}{n_i^2}.$$

Assim, temos o estimador seguinte para a função se sobrevivência, baseado no estimador de Nelson-Aalen para a taxa de falha acumulada e na relação (1.12).

Definição 2.13 (Estimador de Nelson-Aalen).
Sejam para $i = 1, 2, \ldots, n$

- t_i os instantes em que foram observadas falhas na amostra e
- n_i os número de sujeitos em risco no instante t_i;
- f_i o número de falhas ocorridas no intervalo $[t_i, t_{i+1}]$;

Então, o estimador para a função de sobrevivência com base no **estimador de Nelson-Aalen**[3] $\widetilde{S}$ é dado pela função em escada

$$\widetilde{S}(t) = e^{-\sum_{i:t<t_i} \frac{f_i}{n_i}}. \tag{2.5}$$

Nota 2.14 (Estimador de Fleming-Harrington). Na realidade, o estimador (2.5) é conhecido em alguma bibliografia por Fleming-Harrington [6], sendo a designação de estimador de Nelson-Aalen apenas usado para o estimador da taxa de falha acumulada (2.4). Nesse caso, o estimador de Nelson-Aalen pode ser obtido a partir do de Fleming-Harrington através da aplicação de $\hat{\Lambda} = -\ln \widetilde{S}$

Consegue mostrar-se [7] que a variância assintótica deste segundo estimador é dada por

$$\widehat{\mathrm{Var}}\left(\widetilde{S}(t)\right) = \widetilde{S}(t)^2 \widehat{\mathrm{Var}}\left(\hat{\Lambda}\right)(t) = \widetilde{S}(t)^2 \sum_{i:t<t_i} \frac{f_i}{n_i^2}.$$

Exercício R 2.15. Determine a estimativa para a função de sobrevivência com base no estimador de Nelson-Aalen para o exemplo 2.10.

[3] Denotamos o estimador de Nelson-Aalen para a função de sobrevivência por $\widetilde{S}$, por forma a distinguí-lo do estimador de Kaplan-Meier $\hat{S}$.

Resposta.

O estimador de Nelson-Aalen pode ser obtido através dos comandos seguintes (ver figura 2.4):

```
ena<-survfit(Surv(t,cens)~gr,type="fh2",conf.type="plain",
        error="t")
summary(ena)
plot(ena, lty=c(1,2,3), xlab="Tempo", ylab="S(t) por N-A")
legend(1,0.3,lty=c(1,2,3),c("Controlo","Tratamento 1",
        "Tratamento 2"),bty="n")
```

■

Se a partir da função de sobrevivência quisermos obter o estimador de Nelson-Aalen para a taxa de falha acumulada (2.4), podemos fazê-lo aplicando $-\ln$ à função de sobrevivência dada pelo comando `ena$surv`.

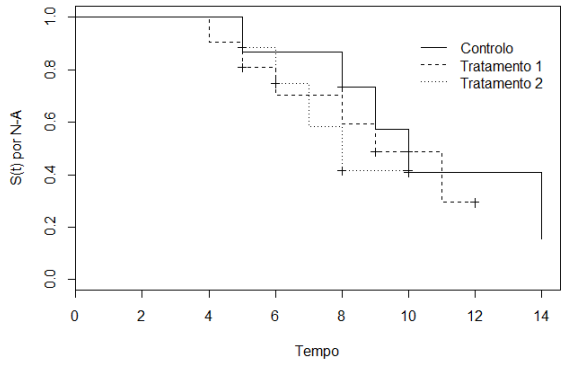


Figura 2.4: Estimativa de Kaplan Meier para a função de sobrevivência do exemplo 2.2.

Os estimadores de Kaplan-Meier e de Nelson-Aalen para a função de sobrevivência dão geralmente estimativas muito próximas, como se pode verificar por comparação das figuras 2.3 e 2.4. Ambos geram estimativas que são funções em escada nos mesmos intervalos, que são definidos pelos instantes de falha. Consegue-se também mostrar [8] que a estimativa de Nelson-Aalen é sempre maior ou igual a de Kaplan-Meier.

Como é visível, nestes modelos não paramétricos só é permitido o estudo da influência de covariáveis categórias. Sempre que existir uma covariável contínua, esta deve ser considerada por classes, por forma a permitir considerar e comparar as estimativas obtidas em cada classe. Para a incorporação de covariáveis contínuas precisamos de outro tipo de métodos, nomeadamente paramétricos ou semiparamétricos. Trataremos deste assunto nos próximos capítulos, mas para já fica um exercício de resumo dos conteúdos até agora abordados.

Exercício 2.16. Considere os dados na tabela seguinte, com dados de pacientes de leucemia. A tabela representa o tempo de remissão (em meses) até nova reincidência, num grupo de controlo e num grupo com tratamento. As censuras estão assinaladas com + .

	Tempo de remissão (meses)							
Controlo	1 5 12	1 8 15	2 8 17	2 8 22	3 8 23	4 11	4 11	5 12
Tratamento	6 11+ 25+	6 13 32+	6 16 32+	6+ 17+ 34+	7 19+ 35+	9+ 20+	10 22	10+ 23

(a) Determine a estimativa de Kaplan-Meier para a função de sobrevivência em cada grupo.

(b) Determine os intervalos de confiança a 95% para a mesma.

(c) Determine se existe efeito de grupo, isto é, se existe uma diferença estatisticamente significativa entre os resultados dos dois grupos.

(d) Considerando o grupo de controlo, determine uma estimativa para a vida média residual vmr no instante $t = 20$. Interprete este valor.

(e) Determine outra estimativa para a função de sobrevivência, usando outro estimador (não-paramétrico) à sua escolha.

Resposta.

(a) Assim obtemos as estimativas para o grupo de controlo e para o grupo em tratamento nas tabelas 2.2 e 2.3, respetivamente. Estas são representadas graficamente na figura 2.5.

(b) Obtemos as tabelas 2.4 e 2.5 com os intervalos de confiança pedidos.

(c) Recorrendo ao R, obtemos um *p-value* para o teste de logrank de comparação das duas estimativas de Kaplan-Meier das funções de sobrevivência por grupo de $p = 0.0000417$. Assim, concluímos com uma confiança de 95% que existem diferenças significativas entre as funções de sobrevivência, ou seja, que o tratamento é eficiente..

Tabela 2.2: Estimativa $\hat{S}$ de Kaplan-Meier para a função de sobrevivência S para o grupo de controlo.

f_i	Intervalos	f_i	n_i	$\hat{S}(t_i^+)$
0	$[0, 1[$	0	21	1.000
1	$[1, 2[$	2	21	0.905
2	$[2, 3[$	2	19	0.810
3	$[3, 4[$	1	17	0.762
4	$[4, 5[$	2	16	0.667
5	$[5, 8[$	2	14	0.571
8	$[8, 11[$	4	12	0.381
11	$[11, 12[$	2	8	0.286
12	$[12, 15[$	2	6	0.190
15	$[15, 17[$	1	4	0.143
17	$[17, 22[$	1	3	0.095
22	$[22, 23[$	1	2	0.048
23	$[23, \infty[$	1	1	0.000

Tabela 2.3: Estimativa $\hat{S}$ de Kaplan-Meier para a função de sobrevivência S para o grupo em tratamento.

t_i	Intervalos	f_i	n_i	$\hat{S}(t_i^+)$
0	$[0, 6[$	0	21	1.000
6	$[6, 7[$	3	21	0.857
7	$[7, 10[$	1	17	0.807
10	$[10, 13[$	1	15	0.753
13	$[13, 16[$	1	12	0.690
16	$[16, 22[$	1	11	0.627
22	$[22, 23[$	1	7	0.538
23	$[23, 35[$	1	6	0.448

(d) Pela expressão de vmr temos

$$\mathrm{vmr}(t) = \frac{\int_t^\infty S(u)du}{S(t)} \approx \frac{\int_t^\infty \hat{S}(u)du}{\hat{S}(t)}$$

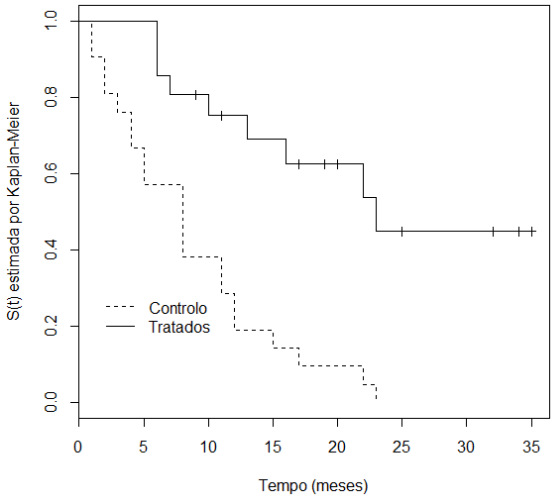


Figura 2.5: Estimativa $\hat{S}$ de Kaplan-Meier para a função de sobrevivência S.

Tabela 2.4: Estimativa para o intervalo de confiança a 95% da função de sobrevivência $\hat{S}$ (baseados na fórmula de Greenwood) para o grupo de controlo.

t_i	Intervalos	f_i	n_i	$\hat{S}(t_i^+)$	$\widehat{\text{Var}}\left(\hat{S}(t_i)\right)$	$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}\left(\hat{S}(t_i)\right)}$
0	$[0,1[$	0	21	1.000	-	-
1	$[1,2[$	2	21	0.905	0.004	[0.779, 1.000]
2	$[2,3[$	2	19	0.810	0.007	[0.642, 0.977]
3	$[3,4[$	1	17	0.762	0.009	[0.580, 0.944]
4	$[4,5[$	2	16	0.667	0.011	[0.465, 0.868]
5	$[5,8[$	2	14	0.571	0.012	[0.360, 0.783]
8	$[8,11[$	4	12	0.381	0.011	[0.173, 0.589]
11	$[11,12[$	2	8	0.286	0.010	[0.092, 0.479]
12	$[12,15[$	2	6	0.190	0.007	[0.023, 0.358]
15	$[15,17[$	1	4	0.143	0.006	[0.000, 0.293]
17	$[17,22[$	1	3	0.095	0.004	[0.000, 0.221]
22	$[22,23[$	1	2	0.048	0.002	[0.000, 0.139]

Em particular no instante $t = 20$, da tabela 2.2 temos $\hat{S}(20) = 0.095$ e

$$\begin{aligned}\int_{20}^{\infty} \hat{S}(u)du &= \int_{20}^{22} \underbrace{\hat{S}(u)}_{0.095} du + \int_{22}^{23} \underbrace{\hat{S}(u)}_{0.048} du \\ &= 2 \times 0.095 + 1 \times 0.048 \\ &= 0.238\end{aligned}$$

Tabela 2.5: Estimativa para o intervalo de confiança a 95% da função de sobrevivência $\hat{S}$ (baseados na fórmula de Greenwood) para o grupo em tratamento.

t_i	Intervalos	f_i	n_i	$\hat{S}(t_i^+)$	$\widehat{\text{Var}}\left(\hat{S}(t_i)\right)$	$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{\text{Var}}\left(\hat{S}(t_i)\right)}$
0	$[0, 6[$	0	21	1.000	-	-
6	$[6, 7[$	3	21	0.857	0.006	[0.707, 1.000]
7	$[7, 10[$	1	17	0.807	0.008	[0.636, 0.977]
10	$[10, 13[$	1	15	0.753	0.009	[0.564, 0.942]
13	$[13, 16[$	1	12	0.690	0.011	[0.481, 0.900]
16	$[16, 22[$	1	11	0.627	0.013	[0.404, 0.851]
22	$[22, 23[$	1	7	0.538	0.016	[0.286, 0.789]
23	$[23, 35[$	1	6	0.448	0.018	[0.184, 0.712]

logo obtemos

$$\text{vmr}(20) \approx \frac{0.238}{0.095} = 2.51.$$

Assim, espera-se que um indivíduo do grupo de controlo que não tenha tido uma remissão até aos 20 meses que a venha a ter (em média) num prazo de dois meses e meio.

(e) Para o estimador de Nelson-Aalen, apresentamos os resultados para o grupo de controlo na tabela 2.6 e para o grupo em tratamento na tabela 2.6. Estas são representadas graficamente na figura 2.6

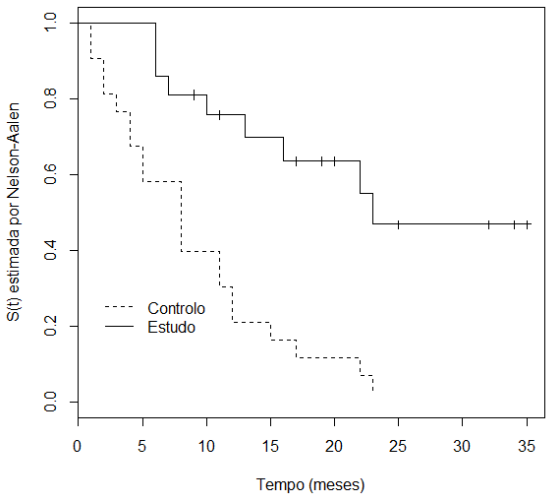


Figura 2.6: Estimativa $\widetilde{S}$ de Nelson-Aalen para a função de sobrevivência S.

Tabela 2.6: Estimativa $\hat{S}$ de Nelson-Aalen para a função de sobrevivência S para o grupo de controle.

t_i	Intervalos	f_i	n_i	$\tilde{\Lambda}(t_i)$	$\tilde{S}(t_i^+)$
0	$[0, 1[$	0	21	0	1
1	$[1, 2[$	2	21	0.095	0.909
2	$[2, 3[$	2	19	0.201	0.818
3	$[3, 4[$	1	17	0.259	0.772
4	$[4, 5[$	2	16	0.384	0.681
5	$[5, 8[$	2	14	0.527	0.590
8	$[8, 11[$	4	12	0.861	0.423
11	$[11, 12[$	2	8	1.111	0.329
12	$[12, 15[$	2	6	1.444	0.236
15	$[15, 17[$	1	4	1.694	0.184
17	$[17, 22[$	1	3	2.027	0.132
22	$[22, 23[$	1	2	2.527	0.080
23	$[23, 23[$	1	1	3.527	0.029

Tabela 2.7: Estimativa $\hat{S}$ de Nelson-Aalen para a função de sobrevivência S para o grupo em tratamento.

t_i	Intervalos	f_i	n_i	$\tilde{\Lambda}(t_i)$	$\tilde{S}(t_i^+)$
0	$[0, 6[$	0	21	0	1
6	$[6, 7[$	3	21	0.143	0.867
7	$[7, 10[$	1	17	0.202	0.817
10	$[10, 13[$	1	15	0.268	0.765
13	$[13, 16[$	1	12	0.352	0.704
16	$[16, 22[$	1	11	0.443	0.642
22	$[22, 23[$	1	7	0.585	0.557
23	$[23, 35[$	1	6	0.752	0.471

■

Capítulo 3

Distribuições Probabilísticas

Neste capítulo analisamos algumas distribuições probabilísticas associadas a modelos paramétricos usuais em Análise de Sobrevivência. A vantagem dos modelos paramétricos é que permitem a introdução de covariáveis contínuas, uma vez que para métodos não paramétricos estas têm de ser tranformadas em categóricas, agregando os dados contínuos em classes. Mais ainda, se o comportamento da variável tempo de falha for conhecido, então o uso de um modelo paramétrico adequado força esse comportamento e logo origina resultados mais precisos. Por outro lado, se esse comportamento for desconhecido, como os modelos paramétricos forçam o comportamento da função a parametrizar, perde-se liberdade na estimação e pode-se introduzir um erro na estimação forçando um comportamento errado. Assim, o uso de modelos paramétricos está sempre dependente da validação do modelo paramétrico para função a estimar, consoante os dados recolhidos.

Como é hábito em Análise de Sobrevivência, os modelos paramétricos aplicam-se à variável tempo de sobrevivência T, isto é, forçam determinado comportamento para a função de sobrevivência dependente da função distribuição de probabilidade $f(t)$ adotado para variável aleatória T e estimando um ou mais parâmetros do modelo paramétrico por forma a adaptar-se aos dados recolhidos. Começamos então a analisar alguns desses modelos, partindo da distribuição de probabilidade associada.

3.1 Distribuições de probabilidade

Enumeramos então algumas das distribuições de probabilidade associadas a modelos paramétricos, partindo dos mais simples para os mais complicados. Esta secção serve principalmente para referência futura, uma vez que muitas destas

distribuições são conhecidas e já suficientemente estudadas. No entanto, tomaremos o contexto da análise de sobrevivência e veremos quais as implicações para o comportamento da taxa de falha e consequentemente da função de sobrevivência de assumirmos que a variável aleatória tempo de falha T tem uma dada distribuição de probabilidade. Um resumo destas implicações e do que deve ser tido em conta para assumirmos como hipótese um dado modelo paramétrico para a função de sobrevivência será abordado na secção 3.1.7

3.1.1 Exponencial

O modelo exponencial é baseado na assunção da distribuição exponencial para variável T, isto é,

$$T \sim \text{Exp}(\lambda).$$

A escolha da notação λ não é inocente, pois como veremos em poucas linhas, o parâmetro λ corresponde à taxa de falha (constante) do modelo exponencial. Pelas propriedades da distribuição exponencial, temos a sua função densidade de probabilidade

$$f(t) = \lambda e^{-\lambda t}, \quad t \geq 0, \tag{3.1}$$

a sua função distribuição de probabilidade,

$$F(t) = 1 - e^{-\lambda t}, \quad t \geq 0, \tag{3.2}$$

e o valor esperado e variância

$$E[T] = \frac{1}{\lambda}, \quad \text{Var}\,(T) = \frac{1}{\lambda^2}.$$

De (1.1), pela relação entre a função de sobrevivência e função distribuição de probabilidade de T, temos que

$$S(t) = 1 - F(t) = e^{-\lambda t}, \quad t \geq 0, \tag{3.3}$$

e logo a função taxa de falha é, por (1.9), constante e dada por

$$\lambda(t) = \lambda, \quad t \geq 0.$$

Assim, para caracterizar o modelo exponencial na ausência de covariáveis[1] basta determinar o valor de λ adequado aos dados recolhidos, assunto que será abordado na secção 3.2. Outra vantagem dos métodos paramétricos é que, como

[1] A introdução de covariáveis em modelos paramétricos será abordada no capítulo 4.

estão já muito estudados, se conhecem algumas medidas sobre a variável T. Por exemplo, do valor esperado de T tira-se diretamente que o tempo médio de vida é

$$t_m = \frac{1}{\lambda}.$$

Por outro lado, como se sabe que para a distribuição exponencial se tem que o percentil p é dado por

$$t_p = -\frac{1}{\lambda}\ln(1-p)$$

pelo que se obtém automaticamente que o tempo mediano de vida é dado por

$$t_{\text{mediano}} = -\frac{1}{\lambda}\ln(0.5) \approx \frac{0.69315}{\lambda}.$$

3.1.2 Weibull

O modelo de Weibull é uma generalização do modelo exponencial com a introdução de mais um parâmetro, uma vez que se a variável T tem distribuição

$$T \sim \text{Wei}(\alpha, \gamma)$$

em que ao parâmetro $\gamma > 0$ se chama o parâmetro de forma e $\alpha > 0$ é o parâmetro de escala, então a sua função densidade de porbabilidade é dada por

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} e^{-\left(\frac{t}{\alpha}\right)^\gamma}, \quad t \geq 0.$$

É assim claro que para $\alpha = 1/\lambda$ e $\gamma = 1$ a expressão anterior se reduz à função densidade de probabilidade (3.1) da distribuição exponencial. Este aspeto fica ainda mais evidente ao considerarmos a função distribuição de probabilidade da Weibull dada por

$$F(t) = 1 - e^{-\left(\frac{t}{\alpha}\right)^\gamma}, \quad t \geq 0.$$

É novamente claro por comparação com (3.2) que para $\gamma = 1$ (e a mudança de variável $\alpha = 1/\lambda$, para uniformizar a notação utilizada) a distribuição de Weibull se reduz à distribuição exponencial. Notamos também que por (1.1), pela relação entre a função de sobrevivência e função distribuição de probabilidade de T, temos que para o modelo Weibull a função de sobrevivência é dada por

$$S(t) = e^{-\left(\frac{t}{\alpha}\right)^\gamma}, \quad t \geq 0, \tag{3.4}$$

e logo a função taxa de falha é, por (1.9), dada por

$$\lambda(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1}, \quad t \geq 0.$$

Para a distribuição Weibull sabemos também que

$$E[T] = \alpha\Gamma\left(1+\frac{1}{\gamma}\right), \quad \mathrm{Var}\,(T) = \alpha^2\left\{\Gamma\left(1+\frac{2}{\gamma}\right) - \Gamma\left(1+\frac{1}{\gamma}\right)^2\right\}, \tag{3.5}$$

em que a função Gama Γ é definida por

$$\Gamma(x) = \int_0^\infty u^{x-1}e^{-u}du \tag{3.6}$$

que satisfaz a propriedade

$$\Gamma(n+1) = n!$$

para n inteiro. Assim, o tempo médio de vida é

$$t_m = \alpha\Gamma\left(1+\frac{1}{\gamma}\right).$$

Mais ainda, como para a distribuição Weibull se tem que o percentil p é dado por

$$t_p = \alpha\left(-\ln(1-p)\right)^{1/\gamma}$$

pelo que se obtém que o tempo mediano de vida é dado por

$$t_{\text{mediano}} = \alpha\left(\ln 2\right)^{1/\gamma}.$$

3.1.3 Log-normal

Outra distribuição usual considerada para a variável tempo de sobrevivência T é a distribuição log-normal

$$T \sim \mathrm{LogNorm}(\mu, \sigma)$$

com parâmetros μ e σ que são, respetivamente, a média e desvio padrão da variável aleatória $X = \ln(T)$. Assim, a sua função densidade de probabilidade é dada por

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma}e^{-\dfrac{(\ln(t)-\mu)^2}{2\sigma^2}}, \quad t > 0,$$

e a função distribuição de probabilidade é dada por

$$F(t) = \Phi\left(\frac{\ln t - \mu}{\sigma}\right), \quad t > 0,$$

em que Φ é a função de distribuição da normal padrão. Assim, temos a função de sobrevivência

$$S(t) = \Phi\left(\frac{\mu - \ln t}{\sigma}\right) \tag{3.7}$$

e a função taxa de falha dada pela relação (1.10)

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

. O valor esperado e a variância são dados respetivamente por

$$E[T] = e^{\mu+\sigma^2/2}, \quad \text{Var}\,(T) = e^{2\mu+\sigma^2}\left(e^{\sigma^2} - 1\right), \tag{3.8}$$

enquanto que o percentil p é definido por

$$t_p = e^{z_p\sigma+\mu}$$

em que z_p é o percentil p da distribuição normal padrão. Assim, o tempo médio de vida é definido por

$$t_m = e^{\mu+\sigma^2/2}$$

enquanto que o tempo mediano mediano de vida é dado por

$$t_{\text{mediano}} = e^{\mu}.$$

3.1.4 Log-logística

A distribuição log-logística é obtida através da distribuição de probabilidade do logaritmo de uma variável aleatória com distribuição logística, ou seja, assumimos que

$$T \sim \text{LogLog}(\alpha, \gamma)$$

se a variável $X = \ln(T)$ tem distribuição logística. A distribuição de probabilidade de uma variável log-logística é dada por

$$f(t) = \frac{\gamma}{\alpha^\gamma}\frac{t^{\gamma-1}}{\left(1 + (t/\alpha)^\gamma\right)^2}, \quad t > 0,$$

e logo a função de distribuição é dada por

$$F(t) = \frac{(t/\alpha)^\gamma}{1 + (t/\alpha)^\gamma}, \quad t > 0.$$

Assim, a função de sobrevivência associada é

$$S(t) = \frac{1}{1 + (t/\alpha)^{\gamma}}, \quad t > 0. \tag{3.9}$$

e a função taxa de falha é

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\gamma}{\alpha^{\gamma}} \frac{t^{\gamma-1}}{1 + (t/\alpha)^{\gamma}}, \quad t > 0.$$

Temos também o valor esperado e a variância são dados respetivamente para $\gamma > 1$ por

$$E[T] = \frac{\alpha\pi}{\gamma \sin(\pi/\gamma)}, \quad \text{Var}(T) = \frac{2\pi\alpha^2}{\gamma \sin(2\pi/\gamma)} - E[T]^2, \tag{3.10}$$

enquanto que o percentil p é definido por

$$t_p = \alpha \left(\frac{p}{1-p} \right)^{1/\gamma}.$$

Assim, para $\gamma > 1$ o tempo médio de vida é definido por

$$t_m = \frac{\alpha\pi}{\gamma \sin(\pi/\gamma)}$$

enquanto que o tempo mediano mediano de vida é dado por

$$t_{\text{mediano}} = \alpha.$$

3.1.5 Gama e Gama Generalizada

A distribuição gama generalizada é também utilizada em várias aplicações de Análise de Sobrevivência, uma vez que esta se adapta a vários comportamentos distintos consoante a escolha dos parâmetros da distribuição, incluindo a distribuição exponencial, de Weibull e Gama. Diz-se que a variável tempo de sobrevivência tem uma distribuição Gama generalizada

$$T \sim \text{GamaGen}(\gamma, \kappa, \alpha)$$

se a função densidade de probabilidade associada for da forma

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma\kappa}} t^{\gamma\kappa-1} e^{-\left(\frac{t}{\alpha}\right)^{\gamma}}, \quad t > 0.$$

em que Γ é a função Gama (3.6). A função distribuição de probabilidade de uma variável Gamma generalizada é dada por

$$F(t) = \frac{g\left(\kappa, (t/\alpha)^{\gamma}\right)}{\Gamma(\kappa)}$$

em que g é a função gama incompleta inferior dada por

$$g(s, x) = \int_0^x t^{s-1} e^{-t} dt.$$

A função gama generalizada é uma distribuição muito geral, pelo que é dificíl indicar propriedades gerais da mesma. Mais ainda, como existem três parâmetros a estimar, são raros os softwares que permitem o uso da distribuição gamma generalizada nas funções de estimação dos pacotes respetivos. De notar no entanto que a distribuição gama generalizada abarca várias das distribuições já abordadas, nomeadamente

- se $\gamma = \kappa = 1$ então $T \sim \mathrm{Exp}(1/\alpha)$;
- se $\kappa = 1$ então $T \sim \mathrm{Wei}(\gamma, \alpha)$;
- se $\kappa \to \infty$ então a distribuição Gama generalizada tende para uma Log-normal;

O caso particular $\gamma = 1$ é também de interesse para nós, uma vez que nesse caso diz-se que T tem uma distribuição gama.

3.1.6 Gama

Tomando $\gamma = 1$ na distribuição gama generalizada, diz-se que T tem uma distribuição gama

$$T \sim \mathrm{Gama}(\kappa, \alpha)$$

com densidade de probabilidade

$$f(t) = \frac{1}{\Gamma(k)\alpha^{\kappa}} t^{\kappa-1} e^{-\left(\frac{t}{\alpha}\right)}, \quad t > 0.$$

e distribuição de probabilidade

$$F(t) = \frac{g\left(\kappa, t/\alpha\right)}{\Gamma(\kappa)}$$

o que se traduz na função de sobrevivência

$$S(t) = \int_0^\infty \frac{1}{\Gamma(k)\alpha^\kappa} u^{\kappa-1} e^{-\left(\frac{u}{\alpha}\right)} du.$$

Para a distribuição gama, já é possível indicar algumas características mais específicas. Por exemplo, a função taxa de falha $\lambda(t) = f(t)/S(t)$ tem um padrão crescente ou decrescente, tendendo para uma constante quando t tende para infinito.

Mais ainda, consegue-se também mostrar que para a distribuição gama:

- Se $0 < \kappa < 1$, a função de densidade f é decrescente e $f(t) \to \infty$ quando $t \to 0$;
- Se $\kappa = 1$, estamos no caso particular da distribuição exponencial;
- Se $\kappa > 1$ a função densidade de probabilidade tem um pico em $t = \alpha(\kappa - 1)$;
- Se $0 < \kappa \leq 1$, f tem concavidade voltada para cima;
- Se $1 < \kappa \leq 2$, f tem concavidade para baixo e depois para cima, com ponto de inflexão em $t = \alpha\left[\kappa - 1 + \sqrt{\kappa - 1}\right]$;
- Se $\kappa > 2$, f tem concavidade voltada para cima, depois para baixo e novamente para cima, com pontos de inflexão em $t = \alpha\left[\kappa - 1 \pm \sqrt{\kappa - 1}\right]$;

Temos também para a distribuição Gama que

$$E[T] = \kappa\alpha, \quad \text{Var}\,(T) = \kappa\alpha^2, \tag{3.11}$$

o que faz com que o tempo médio de vida seja

$$t_m = \kappa\alpha.$$

Para determinar os percentis, não existe fórmula analítica geral, pelo que é preciso utilizar métodos numéricos para obter aproximações dos mesmos.

3.1.7 Escolha do modelo paramétrico apropriado

A escolha de um modelo adequado é fundamental para a validade da estimativa obtida. Assim, nesta secção vamos resumir o comportamento das várias distribuições abordadas e as hipóteses e comportamentos que elas impõe no modelo associado para a função taxa de falha e consequentemente para a função de sobrevivência obtida.

A distribuição exponencial assume que a taxa de falha é constante ao longo tempo. Assim, o modelo exponencial apenas fará sentido em casos em que esta hipótese é válida. Por outras palavras, a distribuição exponencial assume que não há memória sobre o tempo de sobrevivência decorrido, isto é, como a taxa de falha é constante, a probabilidade do sujeito falhar nos próximos instantes é igual caso estejamos perto do tempo inicial ou o sujeito já tenha sobrevivido mil unidades de tempo.

Por outro lado, a distribuição de Weibull é utilizada quando a taxa de falha é monótona, ou seja, conhece-se à partida se esta é crescente ($\gamma > 1$), decrescente ($0 < \gamma < 1$) ou constante ($\gamma = 1$), caso em que se particulariza na distribuição exponencial. Assim, fará sentido utilizar esta distribuição se a ataxa de falha aumentar ao longo do tempo (sendo $\gamma > 1$) ou diminuir ao longo do tempo ($0 < \gamma < 1$).

As distribuições log-normal e log-logística estão associadas a modelos em que as taxas de falha crescem inicialmente, atigem um ponto máximo e depois decrescem. Assim, estes modelos deve ser aplicado para modelar este tipo de fenómenos.

Por outro lado, a distribuição Gama e em especial a distribuição Gama generalizada, conseguem adaptar-se a vários comportamentos diferentes, pelo que são mais gerais. Como desvantagem, devido ao elevado número de parâmetros e à forte não linearidade das funções de probabilidade, geralmente os software de análise de sobrevivência não permitem a utilização direta deste modelo paramétrico.

De referir também que a apresentação das distribuições estatísticas feita nesta secção para o contexto de análise de sobrevivência não é exaustiva. Existem outros modelos que podem ser considerados neste contexto.

3.2 Estimação dos Parâmetros de Modelos

Escolhido o modelo, há que estimar os parâmetros adequados ao conjunto de dados recolhidos, por forma a maximizar a adaptação do modelo aos mesmos. Existem várias formas de o fazer, mas aqui debruçaremo-nos sobre o mais usual: o método da máxima verosimilhança. Entre outras hipóteses está por exemplo o método dos mínimos quadrados. No entanto, o método dos mínimos quadrados não consegue lidar com a presenção de censuras, pelo que a sua utilização no contexto da análise de sobrevivência é limitada à partida.

3.2.1 Método da Máxima Verosimilhança

Caso não haja censuras, o método da máxima verosimilhança passa simplesmente por considerar a função de verosimilhança (sem censuras)

$$\mathcal{L}(t_i, \theta) = \prod_{i=1}^{n} f(t_i, \theta)$$

em que os valores t_i, $i = 1, 2, \ldots n$ são os tempos de falha recolhidos e os valores $\theta = (\theta_1, \theta_2, \ldots, \theta_m)$ são os parâmetros a estimar. Dessa forma, tenta-se escolher os valores dos parâmetros θ que maximizem a função anterior, uma vez que maximizando o produto dos valores $f(t_i, \theta)$ o parâmetro θ maximiza a probabilidade dos tempos t_i recolhidos serem os obtidos por amostragem aleatória da variável com parâmetro θ. Assim, o procedimento passa por considerar o logaritmo da função anterior

$$\mathcal{M}(t_i, \theta) := \ln\left(\mathcal{L}(t_i, \theta)\right) = \sum_{i=1}^{n} \ln\left(f(t_i, \theta)\right)$$

que transforma o produtório num somatório, sem alterar o argumento máximo da função, facilitando os cálculos das derivadas seguintes. De seguida, há que determinar quais os pontos em que a primeira derivada em θ dessa função $\mathcal{M}$ se anula e verificar se a segunda derivada é negativa nesses pontos, por forma a garantir que estamos na presença de um máximo da função de sobrevivência.

No caso de existirem censuras, temos de introduzir essa informação na função de verosimilhança. Se no caso de falha no instante t_i, queremos maximizar a probabilidade desta ocorrer em t_i e logo queremos maximizar $f(t_i, \theta)$ em função de θ para os t_i recolhido, no caso de uma censura em t_i, queremos maximizar a função de sobrevivência em t_i por forma a traduzir a informação de que naquele instante ainda temos informação de existir um sobrevivente (desconhecendo o instante posterior em que este falha). Assim, o método da máxima verosimilhança passa por considerar a função de verosimilhança (com censuras)

$$\mathcal{L}(t_i, \theta) = \prod_{i=1}^{n} \left[f(t_i, \theta)\right]^{\delta_i} . \left[S(t_i, \theta)\right]^{1-\delta_i} \tag{3.12}$$

em que δ_i assume os valores 1 ou 0 consoante o instante t_i é uma falha ou uma censura, respetivamente, conforme definido em (1.14). Da relação $\lambda(t) = f(t)/S(t)$ podemos reescrever a expressão anterior como

$$\mathcal{L}(t_i, \theta) = \prod_{i=1}^{n} \left[\lambda(t_i, \theta)\right]^{\delta_i} .S(t_i, \theta). \tag{3.13}$$

pelo que a função logaritmo da função de verosimilhança é dada por

$$\mathcal{M}(t_i,\theta) = \ln\left(\mathcal{L}(t_i,\theta)\right) = \sum_{i=1}^{n}\left[\delta_i \ln\left(\lambda(t_i,\theta)\right) + \ln\left(S(t_i,\theta)\right)\right].$$

Exercício 3.1. Determine a estimativa de máxima verosimilhança para o parâmetro λ do modelo exponencial.

Resposta.

A partir de (3.13), temos a função de máxima verosimilhança dada no caso exponencial por

$$\mathcal{L}(t_i,\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} e^{-\lambda t_i}.$$

Aplicando o logaritmo, temos

$$\mathcal{M}(t_i,\lambda) = \ln\left(\mathcal{L}(t_i,\lambda)\right) = \sum_{i=1}^{n}\left[\delta_i \ln(\lambda) - \lambda t_i\right] = r\ln(\lambda) - \lambda\sum_{i=1}^{n} t_i$$

em que $r = \sum_{i=1}^{n}\delta_i$ é o número de falhas de entre os n instantes recolhidos (referentes a falhas e censuras). Derivando em λ a expressão anterior, temos

$$\frac{d\mathcal{M}}{d\lambda}(t_i,\lambda) = \frac{r}{\lambda} - \sum_{i=1}^{n} t_i,$$

pelo que igualando a expressão anterior a zero, temos o estimador de máxima verosimilhança para λ dado por

$$\hat{\lambda} = \frac{r}{\sum_{i=1}^{n} t_i}. \tag{3.14}$$

Para verificar que estamos na presença de um máximo, tomamos a segunda derivada, obtendo

$$\frac{d^2\mathcal{M}}{d\lambda^2}(t_i,\lambda) = -\frac{r}{\lambda^2} < 0,$$

pelo que temos que o estimador (3.14) é o estimador de máxima verosimilhança para λ. ■

Infelizmente, para a distribuições mais complexas é difícil obter uma expressão analítica para o estimador de máxima verosimilhança, uma vez que na maior parte dos casos a equação obtida da igualdade da primeira derivada da função de máxima versomilhança a zero não é linear nos parâmetros do modelo. Assim, será sempre recorrer a um método numérico para determinação de zeros da primeira derivada da função, sendo estes geralmente incluídos no software estatístico que permite obter para cada modelo paramétrico considerado, os parâmetros do mesmo.

Exercício R 3.2 (Modelos Paramétricos em R). Considere os dados do exercício 2.2 na página 13. Determine os modelos exponencial, de Weibull, Log-normal e Log-logístico para os dados considerados.

Resposta.

Começamos por carregar o *package* `survival` e os dados com os comandos

```
require(survival)
t<-c(1,3,4,5,5,6,7,8,8,9,10,11,13)
cens<-c(1,0,1,1,0,0,0,1,1,0,1,0,0)
```

De seguida, através dos comandos

```
MExp<-survreg(Surv(t,cens)~1,dist='exponential')
lambda<-1/exp(MExp$coefficients[1])
```

temos que o modelo exponencial (3.3) é obtido pela estimativa de máxima verosimilhança $\hat{\lambda} = 0.06666667$ e dado por

$$\hat{S}_{\text{Exp}}(t) = e^{-0.06666667\,t}.$$

Para o modelo de Weibull (3.4), através dos comandos

```
MWei<-survreg(Surv(t,cens)~1,dist='weibull')
alfaWei<-exp(MWei$coefficients[1])
gamaWei<-1/MWei$scale
```

temos as estimativas de máxima verosimilhança $\hat{\alpha} = 11.73455$ e $\hat{\gamma} = 1.70338$ e logo este é dado por

$$\hat{S}_{\text{Wei}}(t) = e^{-\left(\frac{t}{11.73455}\right)^{1.70338}}.$$

Para o modelo de Log-normal (3.7), temos

```
MLN<-survreg(Surv(t,cens)~1,dist='lognorm')
muLN<-MLN$coefficients[1]
sigmaLN<-MLN$scale
```

temos as estimativas de máxima verosimilhança $\hat{\mu} = 2.291511$ e $\hat{\sigma} = 0.9930993$ e logo este é dado por

$$\hat{S}_{\text{LogNorm}}(t) = \Phi\left(\frac{2.291511 - \ln t}{0.9930993}\right).$$

Finalmente, o modelo de Log-logístico (3.9), é definido pelos comandos

```
MLL<-survreg(Surv(t,cens)~1,dist='loglog')
alfaLL<-exp(MLL$coefficients[1])
gamaLL<-1/MLL$scale
```

de onde obtemos as estimativas de máxima verosimilhança dadas por $\hat{\alpha} = 9.677641$ e $\hat{\gamma} = 1.98028$. Assim, temos o modelo Log-logístico

$$\hat{S}_{\text{LogLog}}(t) = \frac{1}{1 + \left(\frac{t}{9.677641}\right)^{1.98028}}.$$

Por forma a obter os gráficos dos respetivos modelos na figura 3.1, e por forma a podermos comparar com a estimativa não paramétrica de Kaplan-Meier, temos os comandos

```
#KM
ekm<-survfit(Surv(t,cens)~1, conf.int=0)
summary(ekm)
#Graficos
tempos<-seq(0, max(t), length.out = 100)
S_MExp<-exp(-lambda*tempos)
S_MWei<-exp(-(tempos/alfaWei)^gamaWei)
S_MLN<-pnorm((muLN-log(tempos))/sigmaLN)
S_MLL<-1/(1+(tempos/alfaLL)^gamaLL)
plot(ekm, lty=1, xlab="Tempo", ylab="S(t)")
lines(tempos,S_MExp, type="l",lty=2)
lines(tempos,S_MWei, type="l",lty=3)
lines(tempos,S_MLN, type="l",lty=4)
lines(tempos,S_MLL, type="l",lty=5)
legend(1,0.3,lty=c(1,2,3,4,5),c("Kaplan-Meier",
    "Exponencial","Weibull","Log-normal","Log-logistica"))
```

■

Nota 3.3 (Logaritmo da Função de Verosimilhança). Como dado de saída do comando

```
survreg(Surv(t,cens)~1,dist='...')
```

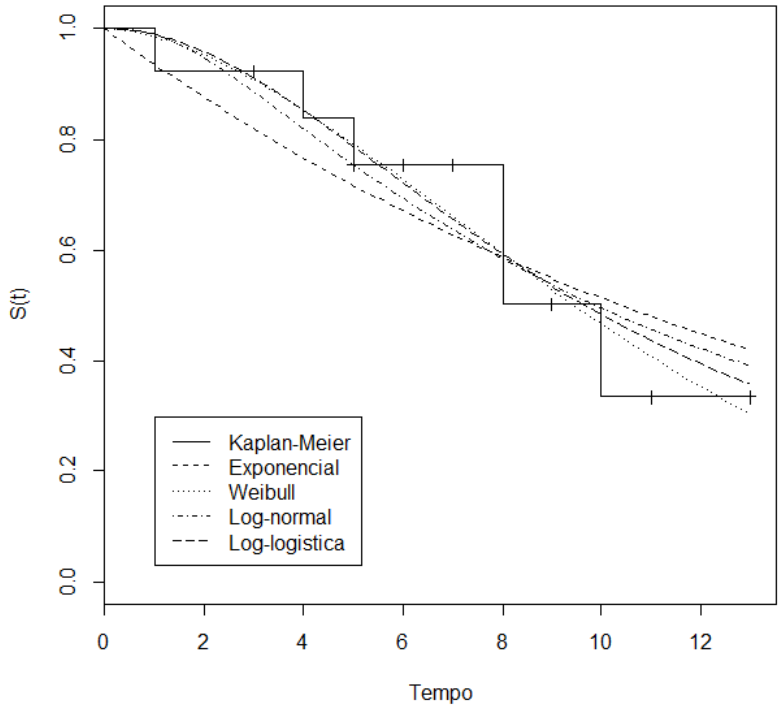


Figura 3.1: Modelos paramétricos para o exercício 3.2.

com a escolha de distribuição pretendida, temos o valor `loglik`, que corresponde ao valor do logaritmo da função de verosimilhança para a estimativa de máxima verosimilhança. Isto quer dizer, portanto, que quanto maior esse valor, menor a verosimilhança. Além disso, esse valor será importante para o teste da razão das verosimilhanças, que estudaremos na seção 3.4.2.

3.3 Adequação do Modelo

A escolha do modelo deve tomar em consideração os aspetos referidos na secção 3.1.7, mas estes não são suficientes para garantir a adequação do método. Na realidade, não basta tomar em consideração o comportamento que se espera *a priori*. É necessário comprovar se o modelo se adapta aos dados recolhidos. Para isso vamos estudar os métodos gráficos para o efeito.

Uma vez que um modelo não paramétrico não impõe condições para a sua aplicabilidade nem para o comportamento da função a estimar, uma forma de verificar a adequação de um modelo paramétrico é comparar graficamente a estimativa não-paramétrica de Kaplan-Meier com a estimativa paramétrica pelo

modelo considerado para a função de sobrevivência. Fazendo então o gráfico dos pontos com abcissa $\hat{S}(t_i)$ e ordenada $\hat{S}_{MP}(t_i)$, em que $\hat{S}$ e $\hat{S}_{MP}$ são respetivamente a estimativa de Kaplan-Meier e pelo modelo paramétrico para a função de sobrevivência no instante de falha ou censura t_i, teremos um modelo adequado se estes pontos estiverem sobre (ou próximos) a reta bissetriz dos quadrantes impares, ou seja, a reta que une a origem do referencial ao ponto $(1, 1)$. Nesse caso, teremos que a estimativa de Kaplan-Meier e do estimador paramétrico são próximas e logo o estimador paramétrico induz o comportamento correto para os dados recolhidos.

Exercício R 3.4 (Validação do modelo paramétrico). Verifique se os modelos paramétricos obtidos no exercício 3.2 são adequados.

Resolução.

Após os comandos em R para a definição dos modelos paramétricos e de Kaplan-Meier na resolução do exercício 3.2, começamos por definir a lista dos instantes de falha ou censura, por

```
tfalhas<-ekm$time
```

De seguida, definimos o valor de a função de sobrevivência estimada por Kaplan-Meier nesses pontos

```
S_KM<-ekm$surv
```

e o valor de cada modelo paramétrico obtido nesses mesmos pontos, através de

```
S_MExp<-exp(-lambda*tfalhas)
S_MWei<-exp(-(tfalhas/alfaWei)^gamaWei)
S_MLN<-pnorm((muLN-log(tfalhas))/sigmaLN)
S_MLL<-1/(1+(tfalhas/alfaLL)^gamaLL)
```

Fazemos então os gráficos respetivos por

```
par(mfrow=c(2,2))
plot(S_KM,S_MExp,xlab="S(t) K-M",ylab="S(t) Exponencial",
        xlim=range(c(0,1)), ylim=range(c(0,1)))
lines(c(0,1),c(0,1), type="l",lty=1)
plot(S_KM,S_MWei,xlab="S(t) K-M",ylab="S(t) Weibull",
        xlim=range(c(0,1)), ylim=range(c(0,1)))
lines(c(0,1),c(0,1), type="l",lty=1)
plot(S_KM,S_MLN,xlab="S(t) K-M",ylab="S(t) Log-Normal",
        xlim=range(c(0,1)), ylim=range(c(0,1)))
lines(c(0,1),c(0,1), type="l",lty=1)
plot(S_KM,S_MLL,xlab="S(t) K-M",ylab="S(t) Log-Logistica",
```

```
        xlim=range(c(0,1)), ylim=range(c(0,1)))
lines(c(0,1),c(0,1), type="l",lty=1)
```

que são apresentados na figura 3.2.

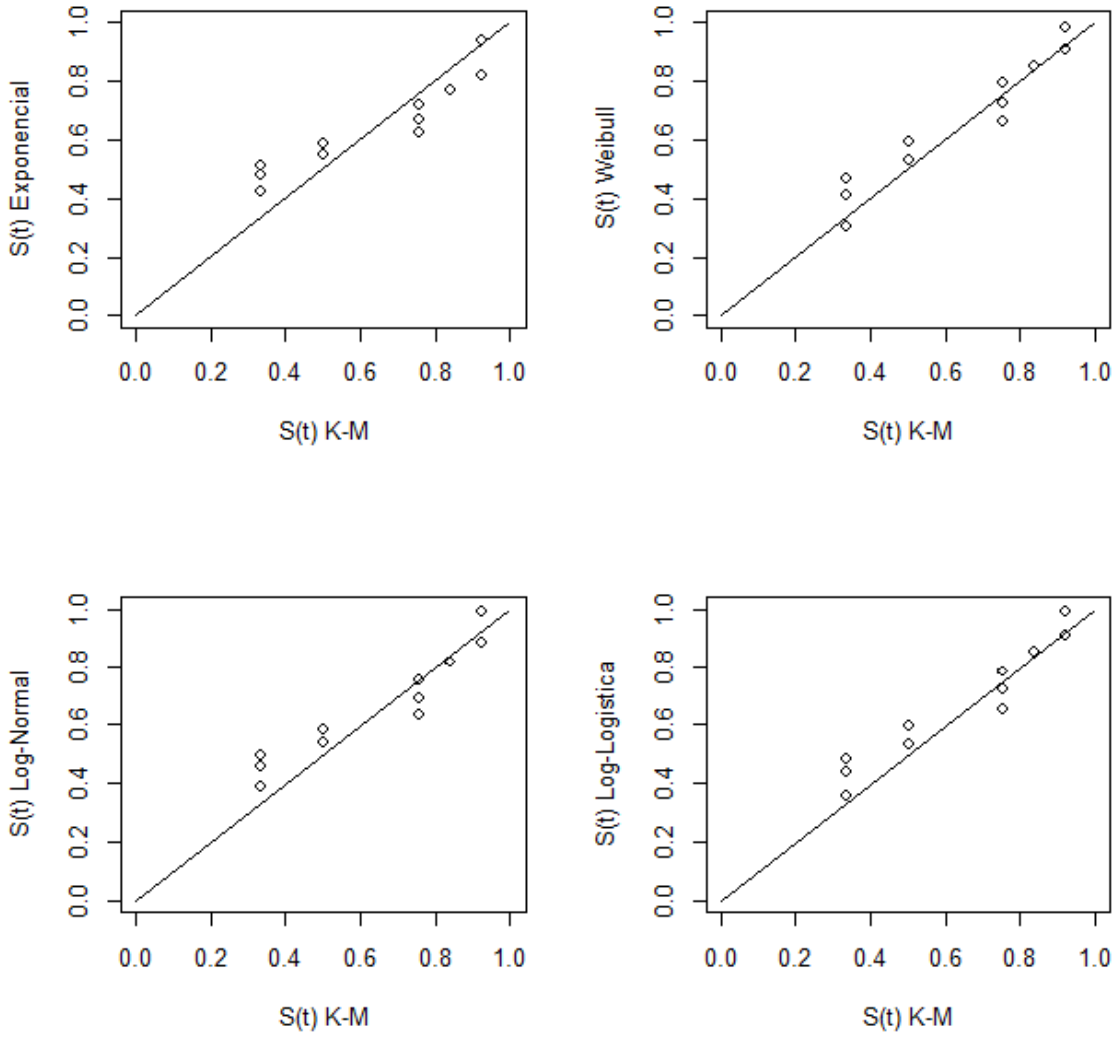


Figura 3.2: Adequação dos modelos paramétricos para o exercício 3.4.

Como conlusão, temos que, à semelhança do que já estava ilustrado na figura 3.1, a figura 3.2 ilustra que todos os modelos têm uma adequação boa, sendo o modelo exponencial aquele com pior adequação. De facto, parece que o modelo exponencial sobreestima a função de sobrevivência quanto esta assume valores inferiores a 0.6 (os pontos estão acima da reta) e subestima quando o valor é superior. De notar também que aparentemente, o modelo de Weibull é o melhor, se

bem que o seu ajuste seja semelhante ao dos restantes modelos. ■

Outra forma de validação passa por linearizar o modelo paramétrico, cuja validade se quer comprovar. Estes partem de obter uma equação da forma

$$mt + b = L(S(t))$$

para cada modelo a partir das suas hipóteses, em que o lado esquerdo representa uma reta e o lado direito é uma função L da função de sobrevivência S. Ao processo anterior chama-se linearização do modelo.. Assim fazendo o gráfico da reta do lado esquerdo e do valor da função do lado direito para os pontos t_i recolhidos, comparamos ambos. Se os pontos estiverem sobre reta considera-se que o modelo é ajustado. De notar que para estimar o valor da função de sobrevivência S, se utiliza o estimador de Kaplan-Meier $\hat{S}$, uma vez que este não apresenta qualquer condição para a sua aplicabilidade nem impõe comportamento predefinidos na função de sobrevivência.

Como exemplo, tomemos o modelo exponencial. A partir de (3.3), ou seja

$$S(t) = e^{-\lambda t},$$

obtemos a linearização do modelo exponencial

$$\lambda t = -\ln(S(t)).$$

Assim, traçando o gráfico da reta com ordenada na origem e declive λ, teremos um bom modelo exponencial se os pontos $-\ln(\hat{S}(t_i))$ estiverem sobre a reta anterior.

Exercício 3.5. Determine uma linerização do modelo

(a) Weibull;

(b) Log-Normal;

(c) Log-Logistico;

Resposta.

(a) A partir de (3.4), ou seja

$$S(t) = e^{-\left(\frac{t}{\alpha}\right)^{\gamma}},$$

pelo que pela aplicação dupla do logaritmo obtemos a linearização do modelo Weibull

$$\left(\frac{t}{\alpha}\right)^{\gamma} = -\ln(S(t)) \Rightarrow \gamma \ln t - \gamma \ln \alpha = \ln\left(-\ln(S(t))\right).$$

Assim, $\ln\left(-\ln(S(t))\right)$ é linear em $\ln t$, pelo que o gráfico de $\ln\left(-\ln(\hat{S}(t))\right)$ com abcissas $\ln t$ deve ser aproximadamente linear. Uma outra hipótese se linearização seria, por exemplo

$$\left(\frac{t}{\alpha}\right)^{\gamma} = -\ln(S(t)) \Rightarrow \frac{t}{\alpha} = \sqrt[\gamma]{-\ln(S(t))}.$$

(b) De (3.7)

$$S(t) = \Phi\left(\frac{\mu - \ln t}{\sigma}\right)$$

temos a linearização do modelo Log-normal

$$\frac{\mu - \ln t}{\sigma} = \Phi^{-1}\left(S(t)\right).$$

Assim, no caso de um modelo log-normal adequado, devem ser aproximadamente colineares os pontos $\left(\ln t_i, \Phi^{-1}\left(\hat{S}(t)\right)\right)$ (em que $\hat{S}$ é a estimativa de Kaplan-Meier e Φ^{-1} são os percentis da distribuição normal standard) sobre uma reta com ordenada na origem $\frac{\mu}{\sigma}$ e declive $-\frac{1}{\sigma}$.

(c) De (3.9)

$$S(t) = \frac{1}{1 + (t/\alpha)^{\gamma}},$$

temos a linearização do modelo Log-logístico

$$1 + \left(\frac{t}{\alpha}\right)^{\gamma} = \frac{1}{S(t)} \Rightarrow \frac{t}{\alpha} = \sqrt[\gamma]{\frac{1}{S(t)} - 1}.$$

pelo que os pontos $\left(t_i, \sqrt[\gamma]{\frac{1}{\hat{S}(t_i)} - 1}\right)$ devem ser aproximadamente colineares sobre uma reta que atravessa a origem e tem declive $\frac{1}{\alpha}$, no caso de um modelo log-logístico adequado.

■

3.4 Testes de hipóteses

Além de uma estimativa pontual, por uma questão de precisão e validade é importante estabelecer intervalos de confiança e/ou testar se um determinado valor que se espera para determinado parâmetro é ou não refutado estatisticamente

pelos dados. Como habitualmente a derivação de intervalos de confiança é consequência de propriedades de grandes amostras e os pormenores técnicos das demonstrações podem ser encontrados em [9].

Sabe-se que para grandes amostras e sob certas condições, pelas propriedades assintóticas do estimador de máxima verosimilhança $\hat{\theta} = \left(\hat{\theta}_1, \hat{\theta}_2, \ldots \hat{\theta}_m\right)$ este tem uma distribuição normal multivariada de média θ e variância definida pela matriz de covariância do estimador. Assim, o intervalo de confiança a $1-\alpha$ de confiança é definido por

$$IC_{1-\alpha} = \left[\hat{\theta} + z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}\left(\hat{\theta}\right)}, \hat{\theta} + z_{1-\alpha/2}\sqrt{\widehat{\mathrm{Var}}\left(\hat{\theta}\right)}\right]$$

Para estimar a matriz de covariância $\mathrm{Var}\left(\hat{\theta}\right)$, sob certas condições, é possivel utilizar por exemplo a aproximação

$$\mathrm{Var}\left(\hat{\theta}\right) \approx -\left[E(\mathcal{D}(\theta))\right]^{-1}$$

em que $\mathcal{D}(\theta)$ é a matriz das segundas derivadas do logaritmo função de verosimilhança $\mathcal{L}$, ou seja, a componente i, j da matriz é definida por

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j}\left(\ln \mathcal{L}(\theta)\right).$$

Em casos em que o valor esperado é desconhecido ou dificil de calcular, pode-se usar $-\left[\mathcal{D}(\theta)\right]^{-1}$, retirando assim o valor esperado. A estimativa de $\mathrm{Var}\left(\hat{\theta}\right)$ é então obtida substituindo na aproximação anterior θ por $\hat{\theta}$.

3.4.1 Teste de Wald

O teste de Wald é uma generalização do teste t-student para testar se o parâmetro θ tem um determinado valor θ_0, nomeadamente considerando o teste de hipóteses

$$\begin{cases} H_0: & \theta = \theta_0, \\ H_1: & \theta \neq \theta_0. \end{cases}$$

A estatística de teste dada por

$$W = (\hat{\theta} - \theta_0)^T \mathcal{D}(\hat{\theta})(\hat{\theta} - \theta_0) \sim \chi^2_m$$

ou seja, tem uma distribuição Qui-quadrado com m graus de liberdade, em que m é o número de parâmetros do modelo. Assim rejeita-se a hipótese nula

para uma significância α se o valor da estatística de teste W é superior a $\chi^2_{m,1-\alpha}$. No caso de um único parâmetro, a estatística de teste reduz-se a

$$W = \frac{(\hat{\theta} - \theta_0)^2}{\widehat{\mathrm{Var}}\left(\hat{\theta}\right)} \sim \chi^2_1.$$

3.4.2 Teste da Razão das Verosimilhanças

Para o teste de hiopóteses

$$\begin{cases} H_0: & \theta = \theta_0, \\ H_1: & \theta \neq \theta_0. \end{cases}$$

o teste da razão das verosimilhanças compara os valores dos logaritmos da função de verosimilhança para o valor estimado $\hat{\theta}$ e o valor de teste θ_0. Assim, sob a hipótese nula, temos a estatística de teste

$$TRV = -2\ln\left(\frac{\mathcal{L}(\hat{\theta})}{\mathcal{L}(\theta_0)}\right) \sim \chi^2_{m,1-\alpha}.$$

Reescrevendo a estatística de teste com a notação $\mathcal{M}(\theta) = \ln(\mathcal{L}(\theta))$ utilizada anteriormente, temos de forma equivalente a estatística de teste

$$TRV = 2\left(\mathcal{M}(\hat{\theta}) - \mathcal{M}(\theta_0)\right) \sim \chi^2_{m,1-\alpha}.$$

A hipótese nula é rejeita se $TRV > \chi^2_{m,1-\alpha}$, sendo que o valor de $\mathcal{M}(\hat{\theta})$ pode ser obtido diretamente no R, conforme indicado na nota 3.3.

O teste de razão de verosimilhanças serve também para estudar se a generalização de modelos é necessária. Por exemplo, dado um conjunto de dados, este teste serve para muito rapidamente verificar se o modelo exponencial é suficiente, ou se a sua generalização pelo modelo de Weibull apresenta uma diferença significativa. Note-se que este teste equivale a testar se $\gamma = 1$.

Exercício R 3.6 (TRV entre modelos generalizados). Indique se para os modelos paramétricos do exercício 3.2, a generalização do modelo exponencial pelo modelo de Weibull é necessária.

Resposta.

Queremos testar a hipótese nula $H_0 : \gamma = 1$ que corresponde ao modelo exponencial, contra a hipótese $H_1 : \gamma \neq 1$ que corresponde à sua generalização pelo modelo Weibull. Assim, após os comandos em R para o cálculo dos modelos exponencial e de Weibull na resolução do exercício 3.2, temos pelos comandos

```
TRV=2*(MWei$loglik[2]-MExp$loglik[2])
qchisq(.95, df=1)
```

que o valor da estatística de teste é $TRV = 1.856335$ e (para uma significância de 5%) o valor de rejeição $\chi^2_{1,0.95} = 3.841459$. Assim, não existe significância estatística para afirmar que o modelo de Weibull é significativamente diferente do modelo exponencial, apesar de (como esperado) tenhamos visto no exercício 3.4, o ajuste pelo modelo de Weibull é melhor que pelo modelo exponencial.

■

Capítulo 4

Modelos de Regressão

Os modelos de regressão permitem a introdução de covariáveis de interesse, que se espera possam ter influência na variável tempo de sobrevivência e consequentemente na função de sobrevivência.

Na realidade, já haviamos visto a aplicabilidade do estudo de covariáveis categóricas para os modelos não-paramétricos. Nesse caso, estimar-se-ia a função de sobrevivência para cada combinação possível de valores das várias covariáveis categóricas, sendo depois possível estudar a influência de cada covariável através do estudo de diferenças significativas entre as funções de sobrevivência para vários valores da mesma, através por exemplo, de um teste de Logrank. No contexto de testes não-paramétricos, para variáveis quantitativas existe a possibilidade de agregar os seus valores em diferentes estratos (ou seja, categorizar a variável quantitativa) e assim optar pelo mesmo procedimento descrito. No entanto, além de se perder informação, é possível que não haja representantes em número suficiente em alguns dos estratos, para que se possam retirar conclusões significativas do ponto de vista estatístico.

A abordagem que trataremos neste capítulo, passa por considerar a introdução de covariáveis (quantitativas ou não) em modelos paramétricos. Para isso, vamos considerar que a alguns parâmetros da distribuição da variável tempo de falha T associada ao modelo são afetados pela covariável de interesse X, estabelecendo-se nessa relação um modelo de correlação. A regressão linear não é geralmente apropriada para o efeito, ou seja, considerar que

$$\theta = \beta_0 + \beta_1 x + \varepsilon$$

em que o erro ε teria uma distribuição normal, não é adequado, uma vez que em geral não se traduz geralmente num modelo válido em análise de sobrevivência. Veremos que este modelo é válido para a regressão da média do logaritmo de

uma variável T com distribuição log-normal, mas nos casos dos modelos exponencial e de Weibull, a regressão linear não será indicada. Ilustramos então os casos mais utilizados neste contexto.

4.1 Regressão Exponencial

No modelo de regressão exponencial sem covariáveis, considera-se a taxa de falha constante. No caso da introdução de covariáveis, consideramos que o parâmetro λ, ou seja, a taxa de falha se mantém constante ao longo do tempo de sobrevivência T, mas depende da covariável X. Dessa forma, assumimos uma regressão exponencial da forma

$$\lambda(t|x) = \varepsilon e^{-\beta_0 - \beta_1 x} \tag{4.1}$$

em que o erro ε tem distribuição exponencial com média unitária. Desta forma, consideramos o estimador de regressão para a taxa de falha dado por

$$\hat{\lambda}(t|x) = e^{-\beta_0 - \beta_1 x}. \tag{4.2}$$

Note-se que podemos linearizar a expressão (4.1) aplicando o logaritmo por

$$Y = \ln(\lambda) = -\beta_0 - \beta_1 x + \nu, \tag{4.3}$$

mas que esta, embora tenha a aparência de uma regressão linear, não o é. Na realidade o erro $\nu = \ln(\varepsilon)$ não tem distribuição normal, como é assumido na regressão linear, mas sim, neste caso uma distribuição de valor extremo padrão.

Assim, substituindo (4.2) na expressão da função de sobrevivência (3.3), temos a expressão

$$S(t|x) = \exp\left(-\frac{t}{e^{\beta_0 + \beta_1 x}}\right), \quad t \geq 0. \tag{4.4}$$

A determinação dos coeficientes β_0 e β_1 define a função de sobrevivência pelo modelo linear, sendo que estes podem ser obtidos pelo método da máxima verosimilhança, tomando em conta as censuras, generalizando o processo descrito na secção 3.2.1 para a otimização dos parâmetros β_0 e β_1.

De notar que no caso com m covariáveis, ou seja, em que se pode considerar $X = (X_1, X_2, \ldots, X_m)$, o caso anterior é generalizado por

$$S(t|x) = \exp\left(-\frac{t}{e^{\beta(x)}}\right), \quad t \geq 0, \tag{4.5}$$

em que

$$\beta(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m \tag{4.6}$$

4.2 Regressão de Weibull

Na regressão de Weibull assumimos que o parâmetro de escala α da expressão (3.4) depende da covariável de interesse X, ou seja, tomamos a regressão exponencial para este parâmetro

$$\alpha = \varepsilon e^{\beta(x)},$$

com a mesma notação para $\beta(x)$ considerada em (4.6) no caso de considerar m covariáveis, de onde temos o estimador para a regressão dado por

$$\hat{\alpha} = e^{\beta(x)}, \tag{4.7}$$

De notar que a partir de (3.4) isto se traduz na função de sobrevivência

$$S(t|x) = \exp\left(-\left(\frac{t}{e^{\beta(x)}}\right)^{\gamma}\right), \quad t \geq 0, \tag{4.8}$$

o que por sua vez se traduz numa taxa de falha dada por

$$\lambda(t|x) = \frac{\gamma}{e^{\gamma\beta(x)}} t^{\gamma-1}, \quad t \geq 0. \tag{4.9}$$

De notar que fixando os valores das covariáveis $X = (X_1, X_2, \ldots, X_m)$, o comportamento da função de sobrevivência é dado pelo comportamento do modelo de Weibull sem covariáveis, estudado na secção 3.1.2.

4.3 Regressão de Log-normal

Para a regressão log-normal, consideramos a regressão linear da média μ da variável $Y = \ln(T)$ em função das covariáveis, isto é,

$$\mu = \beta(x) + \varepsilon, \tag{4.10}$$

em que $\beta(x)$ é definido por (4.6) e ε tem uma distribuição normal com média nula, o que se traduz no modelo de regressão log-normal para a função de sobrevivência dado por

$$S(t|x) = \Phi\left(\frac{\beta(x) - \ln t}{\sigma}\right). \tag{4.11}$$

Assim, assumimos que as covariáveis apenas afetam a média, mas não o desvio padrão.

4.4 Interpretação dos coeficientes de regressão

A interpretação dos coeficientes $\beta_0, \beta_1, \ldots, \beta_m$ do modelo de regressão deve levar em consideração dois aspetos principais. A primeira é que a interpretação dos coeficientes tem de ser cautelosa, uma vez que a regressão usada não é geralmente a linear. A segunda é que a regressão pelos modelos estudados induz uma fator (multiplicativo) na alteração da taxa de falha e não uma alteração aditiva.

Assim, para o modelo de regressão exponencial com uma covariável em que a taxa de falha é dada por (4.2)

$$\lambda(t|x) = e^{-\beta_0 - \beta_1 x} \tag{4.12}$$

temos que se se variar em Δx unidades o valor da covariável x, a taxa de risco do modelo exponencial varia por um fator (multiplicativo) de

$$\frac{\lambda(t|x+\Delta x)}{\lambda(t|x)} = \frac{e^{-\beta_0-\beta_1(x+\Delta x)}}{e^{-\beta_0-\beta_1 x}} = e^{-\beta_1 \Delta x}.$$

Assim, é claro que a taxa de falha não é afetada por um factor linear, mas sim por um fator exponencial em relação à variação da covariável.

Exercício 4.1. Considere o modelo de regressão exponencial com m covariáveis representadas por $X = (X_1, X_2, \ldots, X_m)$. Determine qual o fator por que varia a taxa de falha do modelo com uma varição dos valores das covariáveis, dado respetivamente por $\Delta x = (\Delta x_1, \Delta x_2, \ldots, \Delta x_m)$.

Resposta.

A variação implica uma variação no risco por um fator de

$$\frac{\lambda(t|x+\Delta x)}{\lambda(t|x)} = e^{-\beta_1 \Delta x_1 - \beta_2 \Delta x_2 - \cdots - \beta_m \Delta x_m}.$$

■

Uma abordagem semelhante pode ser feita para o modelo de regressão de Weibull. Assim, sabemos de (4.9) que o risco é afeta por um fator de

$$\frac{\lambda(t|x+\Delta x)}{\lambda(t|x)} = e^{-\gamma(\beta_1 \Delta x_1 + \beta_2 \Delta x_2 + \cdots + \beta_m \Delta x_m)}$$

para uma variação das m covariáveis do modelo de Weibull dada respetivamente por $\Delta x = (\Delta x_1, \Delta x_2, \ldots, \Delta x_m)$.

4.4.1 Variáveis categóricas no modelo de regressão

Outro aspeto importante de referir é como lidar com variáveis qualitativas neste tipo de modelos. Supondo que a variável X_i é categórica e pode assumir, por exemplo os valores $X_i = v_0, v_1, v_2, \ldots, v_n$, em que os valores são qualitativos, não faz sentido considerar uma regressão da forma

$$\theta = \varepsilon e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i+\cdots+\beta_m x_m}$$

para o parâmetro θ, em que x_i assume os valores $0, 1, 2, \ldots, n$ para representar os valores categóricos $v_0, v_1, v_2, \ldots, v_n$, respetivamente, da variável X_i. Na realidade isto implicaria que a diferença entre os valores v_0 e v_1 categóricos (que correspondem aos valores 0 e 1 da variável x_i respetivamente) teria o mesmo significado da diferença entre os valores v_1 e v_2 (que correspondem aos valores 1 e 2 da variável x_i respetivamente), o que não faz sentido para variáveis qualitativas. Geralmente, a diferença entre valores categóricos pode nem estar definida e a estar pode representar diferenças diferentes entres os várioas valores ordenados. Assim, o fator $\beta_i x_i$ da regressão deve ser reestruturado em n parcelas dadas por

$$\sum_{j=1}^{n} \beta_{ij} x_{ij} = \beta_{i1} x_{i1} + \beta_{i2} x_{i2} + \ldots \beta_{in} x_{in}$$

em que as variáveis auxiliares x_{ij} são dicotómicas e assumem apenas os valores 0 ou 1. Nesta representação, para representar o valor v_j da variável X_i com $j \neq 0$ consideramos

$$x_{ij} = 1, \quad x_{ik} = 0, \, \forall k \neq j.$$

e para representar o valor v_0 da variável X_i consideramos todas as variáveis x_{ik} nulas, ou seja,

$$x_{ik} = 0, \, \forall k = 0, 1, \ldots, n.$$

Assim, o coeficiente β_{ij} (ou melhor, o fator $e^{-\beta_{ij}}$) está relacionado com a variação da taxa de falha entre os valores v_j e v_0 da variável X_i, não impondo uma diferença fixa entre valores consecutivos v_j e v_{j+1}. No R, este tipo de análise pode ser feita facilmente introduzindo a covariável com o comando `factor`, indicando assim que se trata de uma variável qualitativa, sendo a regressão considerada como aqui descrito.

4.5 Adequação do modelo

Tal como no caso sem covariáveis abordado na secção 3.3, a adequação do modelo de regressão com covariáveis tem de ser validada. Da mesma forma que no

caso sem covariáveis, iremos ilustrar alguns métodos gráficos para o efeito, comparando as previsões do modelo de regressão com previsões de modelos não-paramétricos. No entanto, e como a introdução direta de covariáveis contínuas nos modelos não-paramétricos não é possível, teremos de adaptar a abordagem.

4.5.1 Resíduos de Cox-Snell

Estes resíduos determinam o ajuste global do modelo e são definidos por

$$\hat{e}_i = \Lambda(t_i|\mathbf{x}_i)$$

em que $\mathbf{x}_i$ é o valor recolhido x para a covariável (se X for uma variável múltipla, então $\mathbf{x}_i$ é um vetor) para o sujeito com evento de falha em t_i. Assim, são a estimativa da função de falha total correspondente ao modelo de regressão obtido que pode ser obtida pela relação (1.13). Temos então os resíduos de Cox-Snell para o modelo exponencial

$$\hat{e}_i = t_i e^{-\beta(\mathbf{x}_i)}, \tag{4.13}$$

para o modelo de Weibull

$$\hat{e}_i = \left(t_i e^{-\beta(\mathbf{x}_i)}\right)^{\gamma} \tag{4.14}$$

e para o modelo log-normal

$$\hat{e}_i = -\ln\left(1 + \Phi\left(\frac{\beta(\mathbf{x}_i) - \ln t}{\sigma}\right)\right). \tag{4.15}$$

com $\beta(x)$ definido como em (4.6). É possível mostrar que os resíduos de Cox-Snell [10, 11, 12] têm uma distribuição exponencial padrão (sob a hipótese do modelo ser adequado) e logo temos que o gráfico dos pontos $\left(\hat{e}_i, \hat{\Lambda}_{\hat{e}_i}(\hat{e}_i)\right)$ devem estar sobre a reta $y = x$. De notar que $\hat{\Lambda}_{\hat{e}_i}$ é a função taxa de falha acumulada para a variável cujos instantes de falha são $\hat{e}_i$, que é estimada pelo estimador de Kaplan-Meier para estes pontos a partir de (1.13), ou seja, por

$$\hat{\Lambda}_{\hat{e}_i}(\hat{e}_i) = -\ln\left(\hat{S}_{\hat{e}_i}(\hat{e}_i)\right).$$

Nos casos dos modelos exponencial ou de Weibull em que existam poucas observações censuradas é conveniente tratar os resíduos censurados como não-censurados, através do ajuste

$$\hat{e}_i = t_i e^{-\beta(\mathbf{x}_i)} + 1, \tag{4.16}$$

ou

$$\hat{e}_i = \left(t_i e^{-\beta(\mathbf{x}_i)}\right)^{\gamma} + 1 \tag{4.17}$$

para os tempos t_i correspondentes a censuras. No caso em que existam muitos dados censurados, esta correção não é válida pelo que a utilização destes resíduos não é válida. De notar que nos casos com muitas censuras os modelos de sobrevivência perdem também fiabilidade, como esperado, pelo que estudar a sua validação é também mais difícil.

4.5.2 Resíduos Padronizados

Outra forma de avaliar a adequação do modelo é a de considerar os resíduos padronizados dados por

$$\hat{\varepsilon}_i = \frac{y_i - \beta(x)}{\hat{\sigma}}$$

em que $y_i = \ln(t_i)$, uma vez que consoante o modelo, se conhece a distribuição do erro ε. Para distribuição exponencial, o erro deve ter uma distribuição de valor extremo padrão (como indicado em (4.3), com ν no lugar de ε), enquanto que por exemplo para a distribuição log-normal estes devem ter uma distribuição normal padrão (como indicado em (4.10)). Esta abordagem acaba assim por ser equivalente e similar à dos resíduos de Cox-Snell.

4.5.3 Resíduos Martingal

Os resíduos de martingal servem para verificar se a regressão da forma (4.6) é adequada, ou seja, se esta deve ser linear, quadrática ou de outra forma. Assim, consideram-se os resíduos martingal

$$\hat{m}_i = \delta_i - \hat{e}_i \tag{4.18}$$

ou seja, a diferença entre a indicação de falha (1.14) e os resíduos de Cox-Snell definidos por (4.13)-(4.15), consoante o modelo de regressão considerado. Grosso modo, os resíduos martingal são indicadores da quantidade de falhas nos dados que não são previstas pelo modelo. Assim, se a nuvem de pontos $(x_{ji}, \hat{m}_i)$ tiver uma tendência linear, a variável X_j está bem representada no modelo. Se por outro lado, a nuvem tiver uma tendência quadrática então a variável X_j deve ser transformada antes de ser introduzida no modelo. O mesmo acontece se o comportamento da variável sofrer alterações a partir de certo valor, uma vez que nesse caso se deve considerar uma discretização da variável em dois estratos (antes e depois do valor), generalizando o processo se existirem mais do que um ponto em que o comportamento é alterado.

4.5.4 Resíduos Deviance

Por fim, os resíduos deviance servem para detetar a presença de outliers no modelo. Como é conhecido, a presença de outliers pode influenciar fortemente o modelo, pelo que deve ser analisado se estes devem ser retirados do modelo. Assim, os resíduos deviance são dados por

$$\hat{d}_i = \mathrm{sgn}(\hat{\mathrm{m}}_\mathrm{i})\sqrt{-2\left(\hat{\mathrm{m}}_\mathrm{i} + \delta_\mathrm{i}\ln\left(\delta_\mathrm{i} - \hat{\mathrm{m}}_\mathrm{i}\right)\right)}. \tag{4.19}$$

em que a função sgn assume o valor -1 ou 1 consoante o argumento é negativo ou positivo, respetivamente. Estes resíduos centram os resíduos martingal em torno de zero, tornando mais simples a identificação de outliers. Os pontos $(t_i, \hat{d}_i)$ indicam a presença de outliers no modelo.

4.6 Significância das Covariáveis

Num modelo de regressão em função de determinadas covariáveis, é importante definir quais as variáveis significativas para o modelo, retirando as restantes. De facto, a presença de covariáveis não significativas para o modelo, além de complicarem o modelo em si, podem causar efeitos na modelação que deveriam ser ignorados.

Uma forma de verificar se a variável X_i é significativa para o modelo é considerar o teste de hipóteses

$$\begin{cases} H_0: & \beta_i = 0 \\ H_1: & \beta_i \neq 0 \end{cases}$$

cuja decisão pode ser encontrada pelo teste de Wald ou pelo teste da razão das verosimilhanças, apresentados nas secções 3.4.1 e 3.4.2, respetivamente.

No entanto a análise não deve parar na análise de significância isolada por covariável. Na realidade deve-se optar por um método complementar que permite a introdução e eliminação faseada de covariáveis, por forma a obter um modelo ótimo. Pode, no conjunto das n covariáveis iniciais, existir alguma que não seja significativa, mas que após a eliminação de algumas outras variáveis (que por estarem correlacionadas com esta) esta adquirir uma significância importante para o modelo. A forma mais comum de fazer esta análise é o método *stepwise*, em duas abordagens que apresentamos de seguida. Este consiste na introdução e eliminação de covariáveis passo a passo, ou seja, de forma faseada.

Por um lado, podemos partir das variáveis significativas da análise independente e eliminar uma a uma cada uma dessas variáveis para testar se o modelo obtido se altera significativamente. No caso em que menos se alterar, retira-se

a covariável correspondente. Continua-se o processo, retirando agora uma a uma as variáveis que permaneceram no modelo e voltando a retirar a que menos influência tem no modelo. Quanto o retirar de qualquer das covariáveis tem alteração significativa no modelo, pára o processo de retirar variáveis. Assim, este processo acaba por eliminar variáveis que possam estar correlacionadas e portanto, com efeito duplicado no modelo. Posteriormente podem-se adicionar uma a uma as variáveis que não foram consideradas significativas na análise independente e adicionar ao modelo a que tiver um efeito mais significativo. O processo é repetido até nenhuma das covariáveis fora do modelo ter um efeito significativo.

Por outro lado, pode-se partir do conjunto total das covariáveis, retirando-as uma a uma e testando qual delas menos afeta o modelo, retirando-a então do modelo adotado. O processo é continuado até que qualquer das covariáveis a retirar tenha um efeito significativo no modelo. Este processo acaba por ser semelhante ao primeiro passo do descrito no parágrafo anterior, partindo apenas de um conjunto maior de covariáveis e sendo por isso, em geral, mais moroso.

Embora seja um processo que conduz geralmente a bons modelos, o método *stepwise* pode ser bastante moroso, especialmente em casos com muitas covariáveis. Nesses casos, a experiência do analista pode ser fundamental para empiricamente tentar eliminar algumas variáveis que sejam esperadas ser menos significativas e mantendo outras que se espera terem mais impacto no modelo. Uma análise de correlação entre as variáveis e uma possível abordagem por análise de componentes principais pode reduzir o elevado número de covariáveis iniciais e assim tornar o processo exequível.

Exemplo 4.2. No ficheiro *gbcs.cvs* (em anexo) está um conjunto de dados relativos a 686 pacientes com cancro na mama. O significado de cada coluna é o seguinte:

Coluna	Variável	Observações
1	Identificação	Número de identificação do paciente
2	Data de Diagnóstico	
3	Data de Recorrência	
4	Data de Óbito	
5	Idade	Idade de Diagnóstico
6	Menopausa	(1=sim, 2=não)
7	Terapia hormonal	(1=sim, 2=não)
8	Tamanho	Tamanho do tumor (mm)
9	Grau	Grau do tumor (1-3)
10	Nódulos	Número de Nódulos (1-51)

11	Progesterona	Número de Recetores de Progesterona (1-2380)
12	Estrogénio	Número de Recetores de Estrogénio (1-1144)
13	Tempo de Remissão	Tempo até nova reincidência (dias)
14	Censura de Remissão	Censura do tempo de remissão (1=falha, 0= censura)
15	Tempo até Óbito	(dias)
16	Censura de Óbito	Censura do tempo até óbito (1=falha, 0= censura)

Pode carregar o ficheiro de dados para o R utilizando comandos da forma seguinte (não esquecer a opção "`header = TRUE`") para algumas variáveis:

```
A<-read.table("C:/.../gbcs.csv",sep=";",header = TRUE);
idade<-A[,5]
tamanho<-A[,8]
grau<-A[,9]
nodulos<-A[,10]
RecProgesterona<-A[,11]
RecEstrogenio<-A[,12]
tempoRemis<-A[,13]
censRemis<-A[,14]
tempoObito<-A[,15]
censObito<-A[,16]
```

Exercício R 4.3. Considere os dados no exemplo 4.2, caracterizados no contexto. Pretende-se estudar se o tamanho do tumor tem influência no tempo de remissão.

(a) Determine o modelo exponencial para o tempo de remissão, considerando o tamanho do tumor como covariável.

(b) Determine o modelo de weibull para o tempo de remissão, considerando o tamanho do tumor como covariável.

(c) Através de um teste de razão das verosimilhanças (TRV) adequado, indique se é necessária a generalização do modelo exponencial pelo modelo de Weibull.

(d) Apresente um gráfico das estimativas do modelo de Weibull para as funções de sobrevivência de pacientes com tamanhos de tumor iguais a 10mm, 50mm e 100mm.

(e) Através de um teste de razão das verosimilhanças (TRV) adequado, determine se a incorporação da covariável "tamanho do tumor"é significativa para o modelo de Weibull.

(f) Através do estudo gráfico de resíduos, comente a adequação do modelo.

Resolução.

(a) Através do *script* seguinte

```
A<-read.table("C:/.../gbcs.csv",sep=";",header = TRUE);
covariavel<-A[,8]
tempos<-A[,13]
cens<-A[,14]
tempos<-tempoRemis
cens<-censRemis
covariavel<-tamanho
# Modelo Exponencial
ajuste<-survreg(Surv(tempos,cens)~covariavel,dist='exponential')
ajuste
beta0<-ajuste$coefficients[1]
beta1<-ajuste$coefficients[2]
```

obtemos os parâmetros $\beta_0 = 8.28310745$ e $\beta_1 = -0.01438481$, logo temos a estimativa para a função de sobrevivência

$$S(t|x) = \exp\left(-\frac{t}{\exp(8.28310745 - 0.01438481x)}\right).$$

(b) Acrescentando ao *script* anterior os comandos

```
# Modelo Weibull
ajustw<-survreg(Surv(tempos,cens)~covariavel,dist='weibull')
ajustw
beta0w<-ajustw$coefficients[1]
beta1w<-ajustw$coefficients[2]
gammaw<-1/ajustw$scale
```

obtemos a estimativa para a função de sobrevivência

$$S(t|x) = \exp\left(-\left[\frac{t}{\exp(8.07018664 - 0.01191895x)}\right]^{1.28703}\right).$$

(c) Pretendemos testar se a generalização do modelo exponencial pelo modelo de Weibull (isto é a introdução do parâmetro γ) melhora substancialmente o valor da função de verosimilhança. Como introduzimos apenas um parâmetro, temos a estatística de teste

$$TRV = 2\left(\log(L_{\mathrm{w}}(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma})) - \log(L_{\mathrm{e}}(\hat{\beta}_0, \hat{\beta}_1))\right) \sim \chi^2_1.$$

Assim, com o *script* seguinte

```
TRV<-2*(ajustw$loglik[2]-ajuste$loglik[2])
pvalue<-1-pchisq(TRV,1)
pvalue
```

obtemos um valor $p = 1.408295 \times 10^{-6}$, logo rejeitamos a hipótese nula (a uma significância de 5%). Assim, a generalização pelo modelo de Weibull é significativa.

(d) Com o *script*

```
time<-0:max(tempos)
SExp<-exp(-(time/exp(beta0w+beta1w*10))^gammaw)
plot(time,SExp,lty=1,type="l",xlim=range(c(0,max(tempos))),
      ylim=range(c(0,1)),xlab="Tempo (dias)",ylab="S(t|x)")
SExp<-exp(-(time/exp(beta0w+beta1w*50))^gammaw)
lines(c(0,time),c(1,SExp),lty=2)
SExp<-exp(-(time/exp(beta0w+beta1w*100))^gammaw)
lines(c(0,time),c(1,SExp),lty=3)
legend(10,0.3,lty=c(1,2,3), c("Tamanho=10mm",
      "Tamanho=50mm","Tamanho=100mm"))
```

temos o gráfico da figura 4.1.

Pelo gráfico parece que quanto maior o tamanho do tumor, menor o tempo de remissão.

(e) Nesta alínea vamos testar se o tamanho do tumor tem influência significativa no tempo de remissão. Para isso, temos a estatística de teste

$$TRV = 2\left(\log(L_{\mathrm{w}}(\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma})) - \log(L_{\mathrm{w}}(\alpha, \gamma))\right) \sim \chi^2_1.$$

Assim, pelo *script*

```
ajustw1<-survreg(Surv(tempos,cens)~1,dist='weibull')
TRV<-2*(ajustw$loglik[2]-ajustw1$loglik[2]) #weibull
pvaluebeta1<-1-pchisq(TRV,1)
```

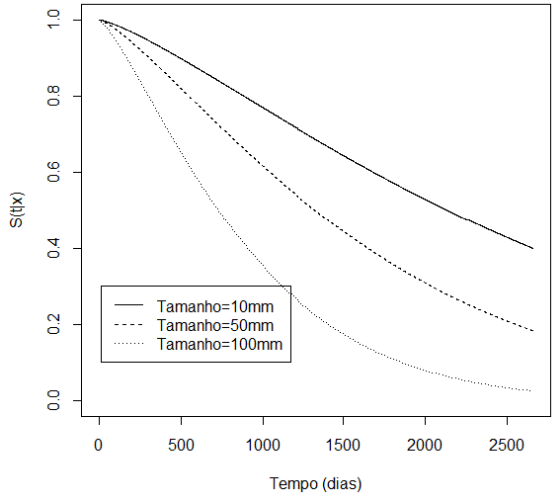


Figura 4.1: Estimativa pelo modelo de Weibull da função de sobrevivência para o tempo de remissão em função do tamanho do tumor.

`pvaluebeta1`
temos um valor $p = 3.489076 \times 10^{-5}$, logo o tamanho do tumor tem influência significativa na função de sobrevivência. O mesmo valor p pode ser recolhido do output do R relativo ao modelo de weibull com uma covariável.

(f) Através do *script*

```
# Residuos Cox-Snell
ei<-(tempos*exp(-(beta0w+beta1w*covariavel)))^gammaw
ei[cens==0]<-ei[cens==0]+1
ekmei<-survfit(Surv(ei,rep(1,length(ei)))~1,
         conf.type="none")
ajustei<-survreg(Surv(ei,rep(1,length(ei)))~1,
         dist='exponential')
alphaei<-exp(ajustei$coefficients[1])
time<-ekmei$time
SExpei<-exp(-time/alphaei)
# Gráficos Residuos Cox-Snell
plot(ekmei, xlab="Tempo", ylab="Estimativa da Função de
Sobrevivência")
lines(c(0,time),c(1,SExpei),lty=2)
legend(2.5,0.8,lty=c(1,2), c("Kaplan-Meier","Modelo"))
```

```
plot(ekmei$surv,SExpei,pch=16, xlab="Kaplan-Meier",
      ylab="Modelo",xlim=range(c(0,1)),ylim=range(c(0,1)))
lines(c(0,1),c(0,1))
# Residuos martingal
ei<-(tempos*exp(-(beta0w+beta1w*covariavel)))^gammaw
mi<-cens-ei
plot(covariavel,mi,pch=16,xlab="Covariavel",
      ylab="R.Martingal")
res<-lm(mi~covariavel)
abline(res)
# Residuos deviance
di<-sign(mi)*sqrt(-2*(mi+cens*log(cens-mi)))
plot(covariavel,di,pch=16,xlab="Covariavel",
  ylab="Residuos deviance",ylim=range(c(min(di),max(di))))
lines(c(0,200),c(0,0))
lines(c(0,200),c(1,1),lty=2)
lines(c(0,200),c(-1,-1),lty=2)
lines(c(0,200),c(2,2),lty=3)
lines(c(0,200),c(-2,-2),lty=3)
```

temos os gráficos da figura 4.2. Pela análise dos resíduos de Cox-Snell, temos um bom modelo até aos desvios da ordem de 0.5, mas a partir daí o modelo parece não se adaptar muito bem. No entanto, os resíduos martingal e deviance, não dão grandes indicações se deveremos usar um modelo de ordem superior (quadrático, cúbico, etc). Isto leva a dizer que, apesar do tamanho ter influência no tempo de remissão, são precisas mais covariáveis para melhor explicar a função de falha, neste caso. No entanto, o facto de o número de censuras ser elevado (387 em 686 casos) tira robustez à análise feita.

■

Exercício R 4.4. Considere os dados no exemplo 4.2, caracterizados no contexto. Pretende-se estudar se o número de receptores de Progesterona e Estrogénio têm influência no tempo até ao óbito.

(a) Determine a estimativa para a função de sobrevivência considerando o modelo lognormal com as covariáveis número de receptores de Progesterona (x_1) e Estrogénio (x_2) da forma

$$\hat{S}(t|x_1,x_2) = \Phi\left(\frac{-\log(t)+\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2}{\hat{\sigma}}\right)$$

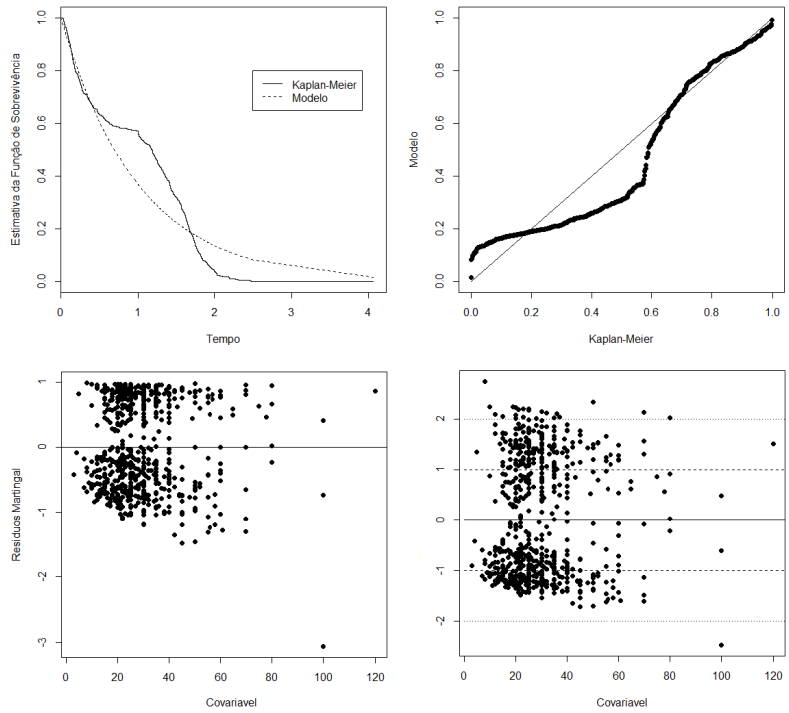


Figura 4.2: Gráficos dos Resíduos. Em cima: Comparação entre a estimativa por kaplan-Meier e o modelo exponencial para os resíduos de Cox-Snell. Em baixo: Gráfico dos resíduos martingal (esquerda) e deviance (direita) em função do valor da covariável tamanho.

utilizando para isso instruções em R da forma:

```
tempos<-tempoObito
cens<-censObito
cov<-cbind(RecProgesterona,RecEstrogenio)
ajuste<-survreg(Surv(tempos,cens)~cov,dist='lognormal')
```

(b) Apresente um gráfico da estimativa da função de sobrevivência pelo modelo para o par de valores $(20, 10)$ correspondente ao número receptores de Progesterona e Estrogénio, respetivamente.

(c) Através de um teste de razão das verosimilhanças (TRV) adequado, indique se a incorporação no modelo do conjunto das duas covariáveis é necessária.

(d) Faça uma análise da adequação do modelo, recorrendo ao estudo dos resí-

duos de Cox-Snell.

(e) Suponha que apenas quer considerar uma covariável no modelo lognormal. De entre as covariáveis "Nódulos", "Progesterona"e "Estrogénio", qual escolheria? Justifique adequadamente.

Resolução.

(a) Acrescentado ao script no exemplo 4.2 os comandos

```
tempos<-tempoObito
cens<-censObito
cov<-cbind(RecProgesterona,RecEstrogenio)
ajuste<-survreg(Surv(tempos,cens)~cov,dist='lognormal')
beta0<-ajust$coefficients[1]
beta1<-ajust$coefficients[2]
beta2<-ajust$coefficients[3]
sigma<-ajuste$scale
```

temos o modelo para a função de sobrevivência em fiunção no número de rectores de Progesterona (x_1) e Estrogénio (x_2)

$$\hat{S}(t|x_1,x_2) = \Phi\left(\frac{-\log(t) + 7.6737013377 + 0.0037702313x_1 + 0.0003132035x_2}{1.004703}\right).$$

(b) Com o *script*

```
time<-0:max(tempos);
Slog<-pnorm((-log(time)+beta0+beta1*20+beta2*10)/sigma)
plot(time,Slog,xlab="Tempo (dias)",ylab="Função S",type="l")
```

obtemos o gráfico da figura 4.3.

(c) Como estamos a testar a significância da introdução de dois parâmetros, temos a estatística de teste

$$TRV = 2\left(\log(L_{\text{l-n}}(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_1,\hat{\sigma})) - \log(L_{\text{l-n}}(\mu,\hat{\sigma}))\right) \sim \chi^2_2.$$

Com o *script*

```
#Ajust Modelo beta=0?
ajustl1<-survreg(Surv(tempos,cens)~1,dist='lognormal')
TRV<-2*(ajust$loglik[2]-ajustl1$loglik[2])
pvaluebeta<-1-pchisq(TRV,2))
pvaluebeta
```

temos o valor $p = 1.706413 \times 10^{-13}$, pelo que a introdução do par de covariáveis é significativo.

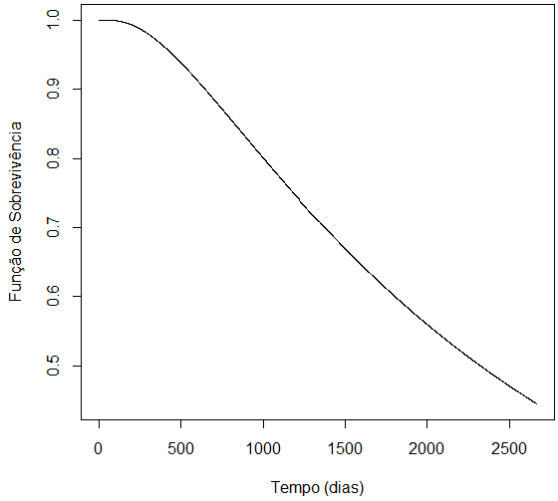


Figura 4.3: Estimativa pelo modelo lognormal da função de sobrevivência para o tempo de óbito com os valores de número de receptores da alínea 2*b*).

(d) Pelo *script*

```
betax=beta0+beta1*covariavel[,1]+beta2*covariavel[,2]
ei<- -log(1-pnorm((log(tempos)-betax)/sigma))
ei[cens==0]<-ei[cens==0]+1
ekmei<-survfit(Surv(ei,rep(1,length(ei)))~1,
       conf.type="none")
ajustei<-survreg(Surv(ei,rep(1,length(ei)))~1,
       dist='exponential')
alphaei<-exp(ajustei$coefficients[1])
time<-ekmei$time
SExpei<-exp(-time/alphaei)
# Gráficos Residuos Cox-Snell
plot(ekmei, xlab="Tempo", ylab="Estimativa da Função
       de Sobrevivência")
lines(c(0,time),c(1,SExpei),lty=2)
legend(2.5,0.8,lty=c(1,2), c("Kaplan-Meier","Modelo"))
#
plot(ekmei$surv,SExpei,pch=16, xlab="Kaplan-Meier",
  ylab="Modelo", xlim=range(c(0,1)),ylim=range(c(0,1)))
lines(c(0,1),c(0,1))
```

obtemos os gráficos da figura 4.4.

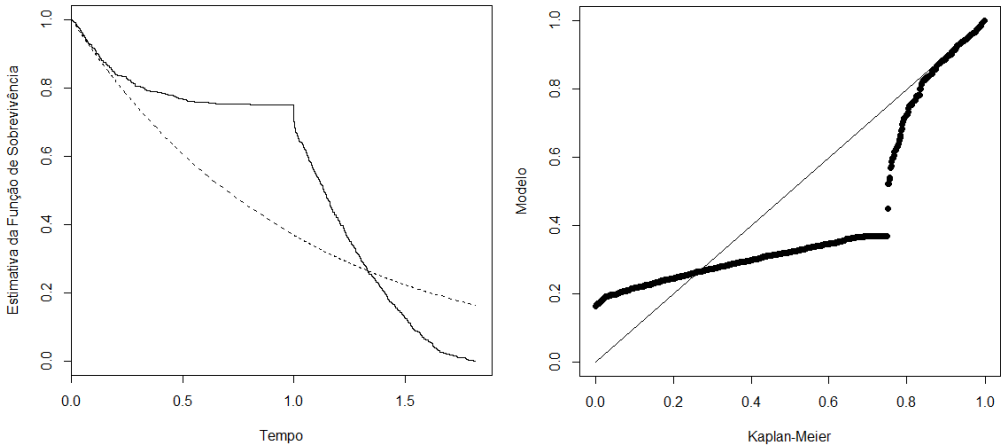


Figura 4.4: Comparação ente a estimativa por Kaplan-Meier e o modelo exponencial para os resíduos de Cox-Snell.

Pela análise do gráfico vemos que o ajuste do modelo é muito mau, pois a estimativa de Kaplan-Meier e pelo modelo exponencial para os resíduos de Cox-Snell são muito diferentes.

(e) Vamos introduzir cada uma das covariáveis de cada vez e verificar qual tem melhor função de máxima verosimilhança. Assim pelo *script*

```
ajust1<-survreg(Surv(tempos,cens)~nodulos,dist='lognormal')
ajust2<-survreg(Surv(tempos,cens)~RecProgesterona,
                                  dist='lognormal')
ajust3<-survreg(Surv(tempos,cens)~RecEstrogenio,
                                  dist='lognormal')
c(ajust1$loglik[2],ajust2$loglik[2],ajust3$loglik[2])
```

temos os valores dos logaritmos das funções de máxima verosimilhança -1590.642, -1584.407, -1609.482, respetivamente. Assim, a variável com mais significado é o número de recetores de Progesterona.

■

Capítulo 5

Modelo de Cox

O modelo de Cox [13] é um modelo semi-paramétrico, ou seja, é um modelo que em parte tem a flexibilidade de um modelo não-paramétrico, mas por outro lado envolve a estimação de alguns parâmetros. Estes parâmetros estão relacionados com hipóteses para o modelo que precisam de ser satisfeitas, sendo um modelo com condições de aplicação, ainda que menos forte que nos modelos paramétricos de regressão usuais.

O modelo de Cox standard parte do princípio dos riscos proporcionais que esclarecemos de seguida.

5.1 Riscos Proporcionais

O modelo de Cox é também designado por modelo de riscos proporcionais, uma vez que assume que a razão entre as taxas de risco entre dois estados de uma covariável é constante. Dito de outra forma, para a covariável X_i, e como habitualmente assumindo regressão exponencial, temos

$$\frac{\lambda(t|x_1, x_2, \ldots, x_i + x, \ldots, x_m)}{\lambda(t|x_1, x_2, \ldots, x_i, \ldots, x_m)} = e^{\beta_i x}. \tag{5.1}$$

Sendo esta a única hipótese a considerar, o modelo de Cox assume que a taxa de falha para os valores $x_1, x_2, \ldots, x_m$ das m covariáveis $X_1, X_2, \ldots, X_m$ é dada por

$$\lambda(t|x_1, x_2, \ldots, x_m) = \lambda_0(t) e^{\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m}. \tag{5.2}$$

em que $\lambda_0(t)$ é denominada a taxa de falha basal e corresponde à taxa de falha quando todas as covariáveis são nulas.

Exercício 5.1. Verifique que a taxa de falha da forma (5.2) satisfaz a condição (5.1) para cada uma das covariáveis $X_1, X_2, \ldots, X_m$.

Assim, a estimativa do modelo de Cox passa por estimar a função $\lambda_0(t)$ (componente não paramétrica) e os parâmetros $\beta_1, \beta_2, \ldots, \beta_m$ (componente paramétrica). A partir da taxa de falha, podemos estimar a função de sobrevivência pelo modelo de Cox a partir da relação

$$S(t|x_1, x_2, \ldots, x_m) = [S_0(t)]^{\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)}. \tag{5.3}$$

em que $S_0(t) = e^{-\int_0^t \lambda_0(u)du}$, ou seja, a função de sobrevivência basal.

É fácil verificar que o modelo de Cox é mais versátil que os modelos de regressão paramétricos estudados no capítulo anterior. Na realidade, enquanto estes assumem o comportamento de λ_0, no modelo de Cox este comportamento é livre *a priori* e estimado *a posteriori* consoante os dados recolhidos. Assim, se o modelo paramétrico for desconhecido, o uso do modelo de Cox é apropriado, o que faz deste modelo muito popular em estudos de sobrevivência. Em particular, a interpretação dos coeficientes paramétricos mantêm-se igual à para modelos paramétricos de regressão discutida na secção 4.4.

5.2 Estimação da componente paramétrica

Partindo da função de verosimilhança (3.13) que recordamos aqui

$$\mathcal{L}(t_i, \theta) = \prod_{i=1}^{n} [\lambda(t_i, \boldsymbol{\beta})]^{\delta_i} . S(t_i, \boldsymbol{\beta}),$$

em que $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$, obtemos por substituição

$$\mathcal{L}(t_i, \theta) = \prod_{i=1}^{n} [\lambda_0(t_i)]^{\delta_i} \, e^{\delta_i(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)} . [S_0(t_i)]^{\exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m)},$$

que é dependente da componente não-paramétrica λ_0. Nesta abordagem, não é possível isolar a estimação dos coeficientes $\beta_1, \beta_2, \ldots, \beta_m$, pelo que optamos por outra abordagem.

Nesse sentido, introduzimos o método da máxima verosimilhança parcial. Este passa por considerar uma probabilidade de falha condicionada. A vantagem de o fazer neste contexto é que como a probabilidade do acontecimento A condicionado a B é dada por

$$P(A|B) = \frac{P(A \cap B}{P(B)}$$

a razão das probabilidades de falha do lado direito anula o elemento λ_0 que é nesta fase desconhecido.

Suponhamos que no instante de falha t_i existem n_i sujeitos em risco dos n sujeitos iniciais e que R_i é o conjunto dos índices dos n_i sujeitos em risco no instante t_i. Cada sujeito é caracterizado pelo conjunto de covariáveis $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(n)}$, em que para cada $j = 1, 2, \ldots, n$ temos que

$$\boldsymbol{x}^{(j)} = (x_1^{(j)}, x_2^{(j)}, \ldots, x_m^{(j)})$$

correspondendo aos m valores das covariáveis consideradas para o sujeito j. Vamos supor sem perda de generalidade que o sujeito j falha em t_i. Então, consideramos a probabilidade do sujeito j falhar em t_i condicionada ao facto de existir uma falha em t_i de entre os que sobreviveram até t_i. Este valor é então dado pela razão entre a taxa de falha para o sujeito j, dada por

$$\lambda(t|\boldsymbol{x}^{(j)}) = \lambda_0(t) e^{\beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \cdots + \beta_m x_m^{(j)}}$$

e a soma das taxas de falhas dos sujeitos em risco no instante t_i dada por

$$\sum_{k \in R_i} \lambda(t|\boldsymbol{x}^{(k)}) = \lambda_0(t) \sum_{k \in R_i} e^{\beta_1 x_1^{(k)} + \beta_2 x_2^{(k)} + \cdots + \beta_m x_m^{(k)}}.$$

Assim, esta probabilidade condicionada no instante t_i é dada por

$$P_i = \frac{\lambda_0(t) e^{\beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \cdots + \beta_m x_m^{(j)}}}{\lambda_0(t) \sum_{k \in R_i} e^{\beta_1 x_1^{(k)} + \beta_2 x_2^{(k)} + \cdots + \beta_m x_m^{(k)}}} = \frac{e^{\beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \cdots + \beta_m x_m^{(j)}}}{\sum_{k \in R_i} e^{\beta_1 x_1^{(k)} + \beta_2 x_2^{(k)} + \cdots + \beta_m x_m^{(k)}}}.$$

Obviamente que em função dos dados obtidos, queremos maximizar o produto de todas as P_i para instantes de falha t_i, pelo que obtemos a função de verosimilhança parcial dada por

$$\mathcal{L}_p(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{e^{\beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \cdots + \beta_m x_m^{(j)}}}{\sum_{k \in R_i} e^{\beta_1 x_1^{(k)} + \beta_2 x_2^{(k)} + \cdots + \beta_m x_m^{(k)}}} \right)^{\delta_i}.$$

A maximização da função de verosimilhança parcial passa agora pelo processo usual de igualar a primeira derivada a zero por forma a obter as estimativas para os respetivos coeficientes $\beta_1, \beta_2, \ldots, \beta_m$. No entanto, como a equação a resolver

não é linear nos coeficientes $\boldsymbol{\beta}$, teremos de aplicar um método numérico para aproximar a solução.

De notar que a expressão anterior não contempla a hipótese de haver empates nos instantes t_i, isto é, que ocorram duas ou mais falhas no mesmo instante, ou uma ou mais censuras num instante de falha. Na primeira situação, deve ser considerada uma correção considerada por Breslow e Peto na discussão do artigo [13] e que está geralmente implementada nos softwares de análise de sobrevivência. Na segunda situação, usa-se a convenção de que as censuras ocorreram depois da falha, pelo que os sujeitos censurados são considerados com sobreviventes no instante t_i.

Nota 5.2 (Razões das taxas de falha). Notamos que a razão entre a taxa de falha correspondente a dois estados diferentes $\boldsymbol{x}^{(1)}$ e $\boldsymbol{x}^{(2)}$ das covariáveis é dado por

$$\frac{\lambda\left(t|\boldsymbol{x}^{(1)}\right)}{\lambda\left(t|\boldsymbol{x}^{(2)}\right)} = \frac{\lambda_0(t)e^{\beta_1 x_1^{(1)}+\beta_2 x_2^{(1)}+\cdots+\beta_m x_m^{(1)}}}{\lambda_0(t)e^{\beta_1 x_1^{(2)}+\beta_2 x_2^{(2)}+\cdots+\beta_m x_m^{(2)}}} = e^{\beta_1\left(x_1^{(1)}-x_1^{(2)}\right)+\beta_2\left(x_2^{(1)}-x_2^{(2)}\right)+\cdots+\beta_m\left(x_m^{(1)}-x_m^{(2)}\right)}$$

e logo não depende da taxa de falha basal λ_0. Desta forma se este for o objetivo do estudo, a taxa de falha basal não precisa de ser estimada.

5.3 Estimação da componente não-paramétrica

Nesta fase e estando completa a estimação dos parâmetros $\boldsymbol{\beta}$ conforme a equação anterior, há que estimar a componente não paramétrica correspondente à taxa de falha basal λ_0. Dessa forma, é comum utilizar um processo não paramétrico, uma vez que a taxa de falha basal é dependente da variável tempo t mas não depende de qualquer parâmetro. Da mesma forma, faremos uso dos instantes de falha t_i e do historial dos acontecimentos de falha até t_i, pelo que a função mais indicada a estimar por regressão é a função taxa de falha acumulada basal Λ_0 que se relaciona com a taxa de falha basal λ_0 por (1.11). Assim, é comum usar a estimativa de Breslow, que surgiu na sua discussão do artigo de Cox [13], dada por

$$\hat{\Lambda}_0(t) = \sum_{i:t_i<t} \frac{f_i}{\sum_{j\in R_i} e^{\beta_1 x_1^{(j)}+\beta_2 x_2^{(j)}+\cdots+\beta_m x_m^{(j)}}} \tag{5.4}$$

em que t_i são os instantes de falha, f_i é o número de falhas em t_i e R_i é o conjunto dos índices dos sujeitos sobreviventes em t_i. Note-se que no caso de não existirem covariáveis (ou seja, de considerarmos os vetores $\boldsymbol{x}^{(j)}$ nulos), esta estimativa coincide com a estimativa de Nelson-Aalen (2.4).

Outra alternativa é usar a estimativa de Kalbfleisch e Prentice

$$\hat{\Lambda}_0(t) = \sum_{i:t_i<t} \left(1 - \left(1 - \frac{f_i e^{\beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_m x_m^{(i)}}}{\sum_{j \in R_i} e^{\beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \cdots + \beta_m x_m^{(j)}}} \right)^{-\exp(\beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_m x_m^{(i)})} \right) \tag{5.5}$$

Em qualquer dos casos, a estimativa para a função de sobrevivência sai diretamente da relação (1.12).

5.4 Validação da hipótese de riscos proporcionais

Uma vez que o modelo parte do pressuposto de riscos proporcionais, a aplicação de resíduos de Cox-Snell não é elucidativa. Uma má adequação ilustrada por resíduos de Cox-Snell pode ser influenciada pela falha do pressuposto de riscos proporcionais. Desta forma, os resíduos de Cox-Snell não conseguem distinguir entre a falha do pressuposto de riscos proporcionais e uma má adequação do modelo (apesar dos resíduos proporcionais se verificarem). Assim, antes da aplicação do modelo, deve-se estudar a hipótese de riscos proporcionais, utilizando os resíduos de Schoenfeld, que apresentamos de seguida. Estes são utilizados para aferir apenas a hipótese de riscos proporcionais e não aferem a qualidade de adequação do modelo.

Desta forma, os resíduos padronizados de Schoenfeld são utilizados, grosso modo, para testar a hipótese dos coeficientes $\boldsymbol{\beta}$ serem independentes do tempo, isto é, comparam a hipótese nula $H_0 : \boldsymbol{\beta}(t) = \boldsymbol{\beta}$ contra a hipótese dos coeficientes não serem constantes ao longo do tempo. Se a hipótese nula for considerada válida, então assume-se de 5.1 que as taxas de falha são proporcionais.

O resíduo de Schoenfeld para o i-ésimo sujeito e k-ésima covariável é dado por

$$r_{ik} = x_k^{(i)} - \frac{\sum_{j \in R_i} x_{jk} e^{\beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \cdots + \beta_m x_m^{(j)}}}{\sum_{j \in R_i} e^{\beta_1 x_1^{(j)} + \beta_2 x_2^{(j)} + \cdots + \beta_m x_m^{(j)}}}$$

Como o somatório dos resíduos de Schoenfeld é nula, utiliza-se os resíduos padronizados de Schoenfeld dados por

$$\mathbf{s}_i = \mathcal{I}^{-1} \mathbf{r}_i$$

em que temos o vetor coluna $\mathbf{r}_i = (r_{i1}, r_{i2}, \ldots, r_{im})$ e $\mathcal{I}$ é a matriz de informação observada, isto é, é a matriz hessiana do logaritmo da função de verosimilhança em função dos parâmetros estimados $\boldsymbol{\beta}$.

Nota 5.3 (Resíduos padronizados de Schoenfeld no R). Os resíduos padronizados de Schoenfeld podem ser calculados no R com o comando
`residuals.coxph(modCox,type='scaledsch')`
em que modCox é o modelo de Cox obtido, como veremos nos exemplos adiante.

Assim, a hipótese nula $\boldsymbol{\beta}(t) = \boldsymbol{\beta}$ constante é equivalente ao gráfico de $s_{ik} + \beta_k$ versus t_i ser uma linha horizontal. Dessa forma, pode-se testar se a reta de regressão linear que melhor se ajusta aos pontos $(t_i, s_{ik} + \beta_k)$ tem declive nulo, ou dito de outra forma, se o gráfico de dispersão anterior representa duas variáveis com correlação nula.

5.5 Validação da adequação do modelo

Tal como nos modelos de regressão, e após verificar que a hipótese de taxas de falha proporcionais é válida, utilizamos os mesmos resíduos apresentados na secção 4.5 para estudar a adequação do modelo de Cox.

Assim , temos os resíduos de Cox-Snell

$$\hat{e}_i = \hat{\Lambda}_0(t_i) e^{\beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_m x_m^{(i)}} \tag{5.6}$$

para cada sujeito $i = 1, 2, \ldots, n$ para avaliar a adequação global do modelo. De forma semelhante, utilizamos os resíduos de martingal

$$\hat{m}_i = \delta_i - \hat{\Lambda}_0(t_i) e^{\beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \cdots + \beta_m x_m^{(i)}} = \delta_i - \hat{e}_i \tag{5.7}$$

para estudar a forma funcional das variável, isto é, se as devemos submeter a alguma transformação quadrática, logarítmica, exponencial, raiz quadrada, entre outras, antes de a introduzir no modelo. Finalmente, os resíduos *deviance*

$$\hat{d}_i = \operatorname{sgn}(\hat{\mathrm{m}}_\mathrm{i}) \sqrt{-2\left(\hat{\mathrm{m}}_\mathrm{i} + \delta_\mathrm{i} \ln\left(\delta_\mathrm{i} - \hat{\mathrm{m}}_\mathrm{i}\right)\right)} \tag{5.8}$$

servem para identificar outliers e valores extremos, ou seja, a presença de dados atípicos que possam influenciar o modelo.

Exemplo 5.4. No ficheiro *whasncc1.csv* está um conjunto de dados relativos a 1494 pacientes com doença cardíaca. Os pacientes estão divididos em dois grupos: o grupo de controlo e o grupo de estudo (ao qual foi feito um tratamento inovador). O significado de cada coluna é o seguinte:

Coluna	Variável	Observações
1	CASE	0=Controlo; 1=Estudo
2	T	Tempo (em dias)
3	FSTAT	Censura: 0= Vivo (censura); 1=Morto (falha);
4	AGE	Idade
5	SEX	Género: 0=Masculino; 1=Feminino
6	CHF	Complicações cardíacas após internamento: 0=não; 1=Sim
7	MIORD	Ordem da ocorrência: 1=primeira; 2=ordem superior
8	LENSTAY	Duração de internamento na unidade hospitalar

Exercício R 5.5.

Considere os dados do exemplo 5.4.

(a) Determine que covariáveis são significativas (a 5%).

(b) Comente a adequação do modelo em termos da hipótese de proporcionalidade entre grupos, considerando apenas as variáveis significativas da alínea anterior.

(c) Considere apenas as CASE, AGE e MIORD. Determine os respetivos coeficientes do modelo de Cox.

(d) Compare graficamente as funções de sobrevivência entre o grupo de controlo e de estudo segundo o modelo anterior, para um indivíduo de 50 anos em primeiro internamento por doença cardíaca.

(e) Comente o aumento da taxa de falha com o aumento de 10 anos na idade (à data de internamento).

Resolução.

(a) Com o *script* em R

```
require(survival)
A<-read.table("../whasncc1.csv",sep=";",header = TRUE);
Case<-A[,1]
tempos<-A[,2]
cens<-A[,3]
Idade<-A[,4]
Sexo<-A[,5]
```

```
CHF<-A[,6]
MIORD<-A[,7]
LENSTAY<-A[,8]
fit0<-coxph(Surv(tempos,cens)~Case+Idade+Sexo+CHF+
  MIORD+LENSTAY,method="breslow")
summary(fit0)
```

obtemos a tabela para o teste de Wald

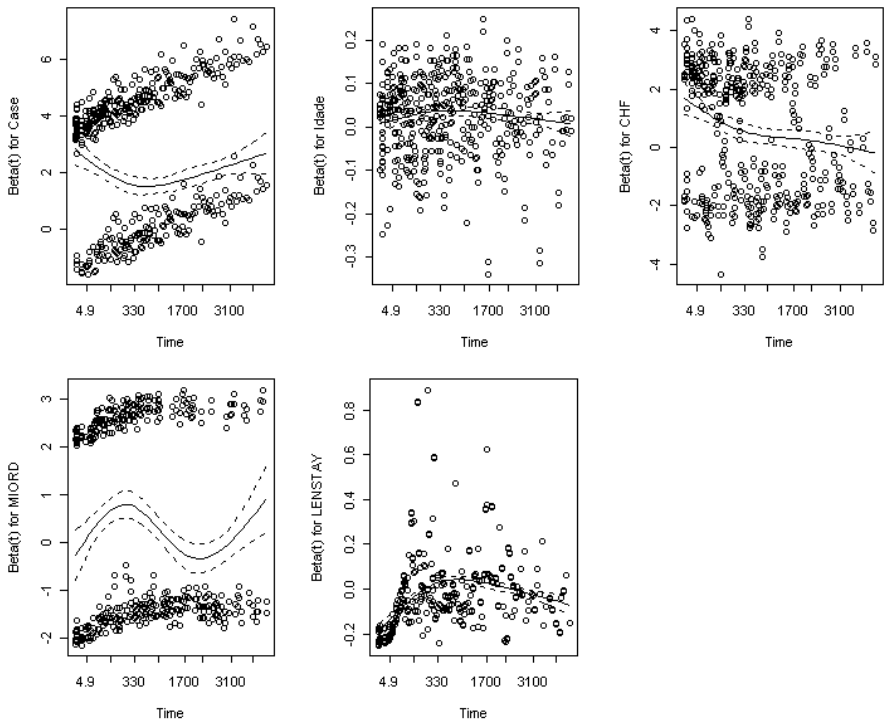
Variável	$\hat{\beta}_i$	p-value
Case	1.962181	$< 2 \times 10^{-16}$
Idade	0.025680	4.41×10^{-12}
Sexo	0.081068	0.33706
CHF	0.548597	2.11×10^{-10}
MIORD	0.230292	0.00476
LENSTAY	-0.020462	7.56×10^{-5}

pelo que apenas a variável Sexo não é significativa.

(b) Com o *script* em R seguinte

```
fit1<-coxph(Surv(tempos,cens)~Case+Idade+CHF+MIORD+
  LENSTAY,method="breslow")
summary(fit1)
resid(fit1,type="scaledsch")
cox.zph(fit1,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit1))
```

obtemos os gráficos para os resíduos de Schoenfeld seguintes:



Os gráficos mostram que a curva adaptada é praticamente constante para a Idade e também pode ser considerada constante para a variável Case. O mesmo pode ser comprovado pelo teste de significância de correlação (obtido pelo mesmo *script*)

Variável	ρ	p-value
Case	0.0481	0.214
Idade	-0.0383	0.294
CHF	-0.1755	9.55×10^{-6}
MIORD	-0.0647	0.106
LENSTAY	0.0701	0.0291

De notar que embora a correlação não seja significativa para a variável MIORD, esta não pode ser considerada como satisfazendo a condição de proporcionalidade, pois a curva não é constante. De facto, a correlação só mede a relação linear, e neste caso a reta que melhor se adapta seria praticamente horizontal. No entanto, a análise do gráfico pode fazer sugerir que é necessário considerar ordens superiores para a variável em causa. Algo

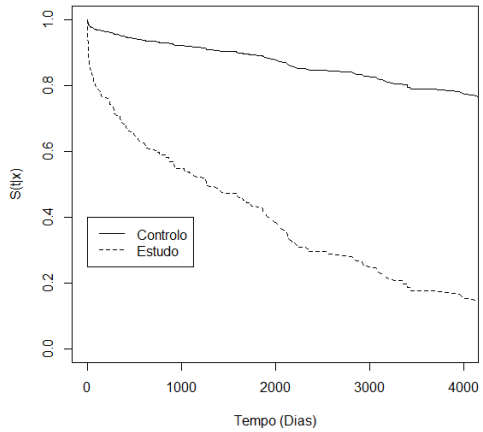
se semelhante se passa com a covariável LENSTAY, uma vez que embora a correlação linear seja significativa e muito próxima de zero, o gráfico indica que esta não é constante.

(c)
```
fit1<-coxph(Surv(tempos,cens)~Case+Idade+MIORD,
  method="breslow")
summary(fitT)
beta1<-fitT$coefficients[1]
beta2<-fitT$coefficients[2]
beta3<-fitT$coefficients[3]
```
obtemos os coeficientes

Variável	$\hat{\beta}_i$
Case	1.98694
Idade	0.03237
MIORD	0.26930

(d) Com o script

```
Ht<-basehaz(fitT,centered=F);
temp<-Ht$time
H0<-Ht$hazard
S0<-exp(-H0)
S1<-S0^exp(beta1*0+beta2*50+beta3*0)
S2<-S0^exp(beta1*1+beta2*50+beta3*0)
plot(temp,S1,lty=1,type="l",xlim=range(c(0,4000)),
    ylim=range(c(0,1)),xlab="Tempo (Dias)",ylab="S(t|x)")
lines(c(0,temp),c(1,S2),lty=2)
legend(1,0.4,lty=c(1,2,3,4,5,6), c("Controlo","Estudo"))
```
obtemos o gráfico seguinte:



Assim, concluimos que o grupo de estudo apresenta uma função de sobrevivência menor.

(e) Temos

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t)e^{\hat{\beta}_1 x_1 + \hat{\beta}_2(x_2+10) + \hat{\beta}_3 x_3}}{\lambda_0(t)e^{\hat{\beta}_1 x_1 + \hat{\beta}_2(x_2) + \hat{\beta}_3 x_3}} = e^{10\hat{\beta}_2} = 1.382182,$$

ou seja, a taxa de falha aumenta por um fator de cerca de 1.4.

■

Exercício R 5.6.

Considere os dados do exemplo 5.4. Pretende-se determinar a interação entre a variável idade e duração de internamento.

(a) Determine os parâmetros (da componente paramétrica) do modelo de Cox considerando apenas as variáveis idade, duração de internamento e a sua interação.

(b) Indique se as variáveis anteriores e a sua interação são significativas para o modelo.

(c) Indique se o princípio da proporcionalidade entre as taxas de falha é satisfeito.

(d) Determine a razão entre as taxas de falha, quando o tempo de internamento aumenta 10 dias para um indivíduo de 60 anos.

(e) Compare as funções de sobrevivência previstas pelo modelo para os pares de idade e duração de internamento seguintes:$(50, 10)$,$(50, 20)$,$(60, 10)$,$(60, 20)$.

Resolução.

(a) Pelo *script* seguinte:

```
fit2<-coxph(Surv(tempos,cens)~Idade*LENSTAY,method="breslow")
summary(fit2)
beta1<-fit2$coefficients[1]
beta2<-fit2$coefficients[2]
beta3<-fit2$coefficients[3]
```

temos os coeficientes do modelo e a sua significância.

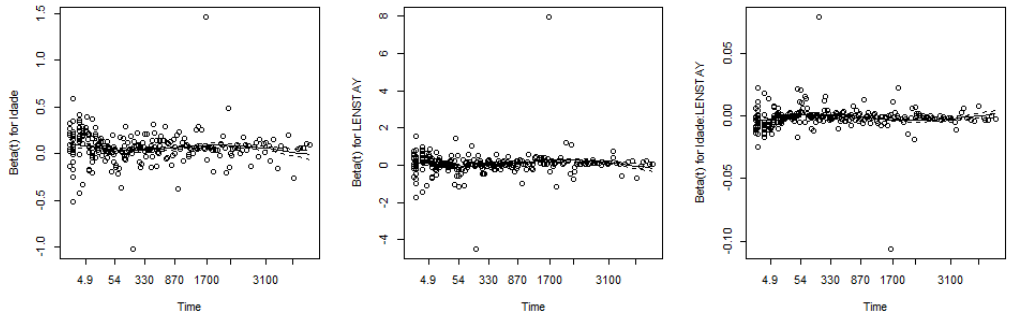
Variável	$\hat{\beta}_i$	p-value
Idade	0.0632741	$< 2 \times 10^{-16}$
Duração de Internamento	0.0903822	4.63×10^{-5}
Idade$\times$Duração de Internamento	-0.0014334	4.70×10^{-6}

(b) Pela tabela anterior, as variáveis e a sua interação são significativas a uma significância de 5%, porque os valores de p para o teste de Wald são inferiores a 0.05.

(c) Pela análise dos resíduos de Schoenfeld dada pelo *script* seguinte

```
resid(fit2,type="scaledsch")
cox.zph(fit2,transform="identity")
par(mfrow=c(2,4))
plot(cox.zph(fit2))
```

obtemos o gráfico



e a tabela de correlação

Variável	ρ	p-value
Idade	-0.0219	0.51712
Duração de Internamento	0.0500	0.10924
Idade $\times$ Duração de Internamento	-0.0451	0.13756

.

Assim, como as curvas adaptadas são praticamente horizontais e os valores de significância das correlações são superiores a 0.05, não existe relevância estatística para rejeitar a hipótese de proporcionalidade entre taxas de falha.

(d) A taxa de falha é dada por

$$\lambda(t) = \lambda_0(t)e^{\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1 x_2}$$

ou seja, obtemos a razão

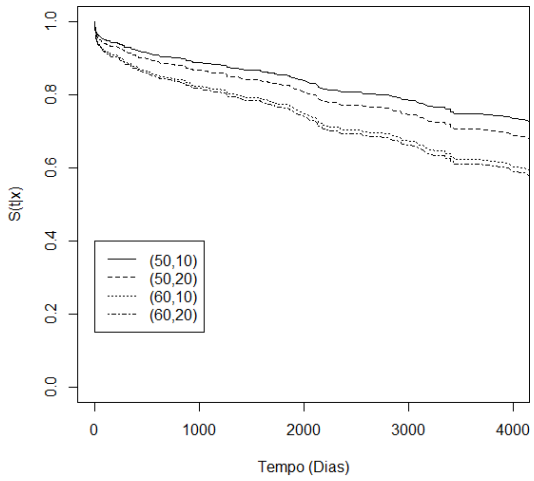
$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t)e^{\hat{\beta}_1 60 + \hat{\beta}_2(x_2+10) + \hat{\beta}_3 60(x_2+10)}}{\lambda_0(t)e^{\hat{\beta}_1 60 + \hat{\beta}_2 x_2 + 60\hat{\beta}_3 x_2}} = e^{10\hat{\beta}_2 + 60\times 10\hat{\beta}_3} = 1.044727,$$

ou seja, não existe uma alteração muito relevante.

(e) Com o *script*

```
Ht<-basehaz(fit2,centered=F)
temp<-Ht$time
H0<-Ht$hazard
S0<-exp(-H0)
S11<-S0^exp(beta1*50+beta2*10+beta3*50*10)
S12<-S0^exp(beta1*50+beta2*20+beta3*50*20)
S21<-S0^exp(beta1*60+beta2*10+beta3*60*10)
S22<-S0^exp(beta1*60+beta2*20+beta3*60*20)
plot(temp,S11,lty=1,type="l",xlim=range(c(0,4000)),
  ylim=range(c(0,1)),xlab="Tempo (Dias)",ylab="S(t|x)")
lines(c(0,temp),c(1,S12),lty=2)
lines(c(0,temp),c(1,S21),lty=3)
lines(c(0,temp),c(1,S22),lty=4)
legend(1,0.4,lty=c(1,2,3,4), c("(50,10)","(50,20)",
  "(60,10)","(60,20)"))
```

obtemos o gráfico seguinte:



Tal como a alínea anterior demonstrou, não existe uma grande diferença entre os pares (60,10) e (60,20). De qualquer das formas, tanto o aumento da idade como o aumento da duração de internamento decrescem a função de sobrevivência.

■

Capítulo 6

O que não foi dito...

Embora não sejam tratadas em detalhe neste texto, vamos enumerar algumas possíveis extensões do que foi aqui apresentado. Estas servirão para quem esteja interessado no assunto tenha algumas orientações para pesquisa futura, para de alguma forma eliminar algumas omissões propositadas durante este percurso.

6.1 Modelos de tempo de vida acelerado

Na apresentação feita nos capítulos 3 a 5, foi sempre considerado que a presença de covariáveis significativas influenciaria a taxa de falha de forma proporcional. No entanto, nem sempre será o caso, na prática. Um outro tipo de modelos que é possível de utilizar é o chamado modelo de tempo de sobrevivência acelerado. Neste caso, a presença de covariáveis acelera ou diminui o tempo de sobrevivência na função de sobrevivência basal, ou seja, considera-se um modelo da forma

$$S(t|\boldsymbol{x}) = S_0(tg(\boldsymbol{x}))$$

em que a função g altera a escala de tempo em função do valor das covariáveis $\boldsymbol{x}$. Mais detalhes deste tipo de modelos podem ser encontrados por exemplo em [12, 14].

6.2 Modelo de Cox estratificado

Quando a hipótese de taxas de falha proporcionais não é verificada para determinada variável, não se pode aplicar o modelo de Cox conforme apresentado no capítulo 5. Nesse caso, uma possibilidade pode ser a de estratificar a variável em causa e encontrar um modelo de Cox em cada estrato. De alguma forma,

a ideia é semelhante à avançada para a aplicação de métodos não paramétricos a variáveis contínuas, sendo que as causas para a estratificação são diferentes. No caso dos métodos não-paramétricos, estratificava-se a variável contínua por forma a obter um modelo não-paramétrico em cada estrato, uma vez que no caso de modelos não paramétricos não é possível introduzir covariáveis diretamente no modelo. No caso do modelo de Cox, a estratificação ocorre para se obter um modelo diferente de Cox em cada estrato e assim eliminar a necessidade de taxas de falha proporcionais entre estratos. Mais detalhes sobre este modelo de Cox estratificado podem ser obtidos, por exemplo, em [12, 14].

6.3 Modelo de Cox com covariáveis dependentes do tempo

O modelo de Cox que ilustrámos neste capítulo assume entre outras hipóteses que as covariáveis não se alteram no tempo. No caso de se alterarem, este não pode ser aplicado. Note-se que a suposição de não variação das covariáveis pode ser limitativa em alguns casos. Por exemplo o estado civil de um sujeito pode-se alterar durante um estudo, assim como por exemplo o seu nível de glicémia em jejum ou o peso do paciente. Nesse contexto já não faz sentido falar-se em risco proporcionais, pois sendo a taxa da falha agora dada por

$$\lambda(t|x_1(t), x_2(t), \ldots, x_i(t), \ldots, x_m(t)) = \lambda_0(t)e^{\beta_1 x_1(t)+\beta_2 x_2(t)+\cdots+\beta_m x_m(t)} \tag{6.1}$$

a razão das taxas de falha entre dois sujeitos 1 e 2 é agora dada por

$$\frac{\lambda(t|x_1^{(1)}(t), x_2^{(1)}(t), \ldots, x_m^{(1)}(t))}{\lambda(t|x_1^{(2)}(t), x_2^{(2)}(t), \ldots, x_m^{(2)}(t))} = e^{\beta_1\left(x_1^{(1)}(t)-x_1^{(2)}(t)\right)+\beta_2\left(x_2^{(1)}(t)-x_2^{(2)}(t)\right)+\cdots+\beta_m\left(x_m^{(1)}(t)-x_m^{(2)}(t)\right)}. \tag{6.2}$$

e logo não é constante ao longo do tempo. Este modelo é portanto ainda mais versátil que o standard e é uma generalização importante do método de Cox. Mais detalhes deste tipo de modelos podem ser encontrados por exemplo em [12, 14].

6.4 Modelos *first hitting time*

Uma nova porta que se vem abrindo para o estudo de sobrevivência é o de assumir que existe um modelo estocástico subjacente à falha. Assim, assume-se que o

sujeito inicia com um valor positivo da variável Y, que corresponde a um estado saudável, e que a falha ocorre quando a variável estocástica associada Y assume o valor nulo. Esta variável estocástica é impossível de medir na realidade, mas assume-se que está subjacente ao processo. Assim, modelar o tempo de falha é equivalente a modelar o processo estocástico da variável Y. Estas técnicas estão associadas a modelos de *first hitting time* (FHT) e regressão *threshold* (TR), que podem ser consultados com mais detalhe em [15, 16].

Índice

Bibliografia

[1] E. L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):pp. 457–481, 1958.

[2] N. Breslow and J. Crowley. A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, 2(3):437–453, 05 1974.

[3] John D. Kalbfleisch and Ross L. Prentice. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, Inc., 2002.

[4] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):pp. 945–966, 1972.

[5] Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):pp. 701–726, 1978.

[6] Thomas R. Fleming and David P. Harrington. *Counting Processes and Survival Analysis*. John Wiley & Sons, Inc., 2005.

[7] Odd O. Aalen and Søren Johansen. An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, 5(3):pp. 141–150, 1978.

[8] George Bohoris. Comparison of the cumulative-hazard and kaplan-meier estimators of the survivor function. *IEEE Transactions on Reliability*, 43(2):pp. 230–232, 1994.

[9] David Roxbee Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman and Hall, 1974.

[10] E. J. Snell D. R. Cox. A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, 30(2):248–275, 1968.

[11] J.F. Lawless. *Statistical models and methods for lifetime data*. John Wiley & Sons, New York, 1982.

[12] Enrico Antonio Colosimo and Suely Ruiz Giolo. *Análise de Sobrevivência Aplicada*. Editora Blücher, 2006.

[13] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[14] David G. Kleinbaum and Michel Klein. *Survival Analysis, a Self-Learning Text*. Springer, 2nd edition edition, 2005.

[15] Mei-Ling Ting Lee. *Lifetime Models and Risk Assessment*. John Wiley & Sons, Ltd, 2014.

[16] D. Stogiannis, C. Caroni, C. E. Anagnostopoulos, and I. K. Toumpoulis. Comparing first hitting time and proportional hazards regression models. *Journal of Applied Statistics*, 38(7):1483–1492, 2011.